DeepSpeed & AnomalyGPT

Project Research Internship

NCHC – 24/25

Pablo Mollá

NARLabs

National Applied Research Laboratories

université PARIS-SACLAY

Presenter

Alicante, Spain
Mathematics in Madrid
Master in Data Science in Paris
Hobbies: Gym, Reading, Traveling

UNIVERSIDAD COMPLUTENSE MADRID

université PARIS-SACLAY

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

EUSPA European Union Agency for the Space Programme

NARLabs 國家實驗研究院
國家高速網路與計算中心
National Center for High-performance Computing

Agenda

- **Motivation**
- **Introduction**
  - DeepSpeed & DeepSpeed-Chat
  - AnomalyGPT
- **Specific Contributions of the Project**
  - Decoder Layer
  - Dataset inclusion: AeBAD
  - Data Augmentation: Poisson Dist. Editing Technique
  - Multi-Node Training Protocol and Implementation
  - Resource Consumption Monitoring
- **Experimental Results**
  - Baseline Model vs 5 New Model Implementations
- **Literature Survey**
  - LVLMs, LLMs and Encoders: ImageBind, PandaGPT, LlaVA, Vicuna, CLIP
  - Distributed Deep Learning Frameworks: DeepSpeed & Horovod
- **Conclusions & Future Research**


Motivation

- Open-Source Deep Learning Library
- AnomalyGPT

deepspeed

DEEPSPEED CHAT

Microsoft

- Speed Up Training Time
- Reduce Computational Costs
- Enhance performance & scalability of AI Models
- Democratization AI via accessible tools

DeepSpeed Adoption

- Megatron-Turing NLG (530B)
- Jurassic-1 (178B)
- BLOOM (176B)
- GLM (130B)
- xTrimoPGLM (100B)
- YaLM (100B)
- GPT-NeoX (20B)
- AlexaTM (20B)
- Turing NLG (17B)
- METRO-LM (5.4B)

LLMs Integration

Lightning

Transformers

Accelerate

mosaicML

MII Engine

Determined AI

Open-Source Deep Learning Frameworks

DeepSpeed Chat

Specialized extension of DeepSpeed framework optimized for RLHF

Easy, **Fast**, **Affordable** Reinforcement Learning with Human Feedback (**RLHF**) training of ChatGPT-like Models **AT ALL SCALES**

Table 1: Single-Node 8x A100: Training Time and Corresponding Approximate Cost on Azure.

GPUs	OPT-6.7B	OPT-13B	OPT-30B	OPT-66B
8x A100-40GB	5.7 hours	10.8 hours	1.85 days	NA
8x A100-80GB	4.1 hours ($132)	9 hours ($290)	18 hours ($580)	2.1 days ($1620)

Table 2: Multi-Node 64x A100-80GB: Training Time and Corresponding Approximate Cost on Azure.

GPUs	OPT-13B	OPT-30B	OPT-66B	OPT-175B
64x A100-80G	1.25 hours ($320)	4 hours ($1024)	7.5 hours ($1920)	20 hours ($5120)

Table 3: Max Model Size Supported by DeepSpeed-HE on a Single GPU.

	V100 32G	A6000 48G	A100 40G	A100 80G
Model Size	OPT-2.7B	OPT-6.7B	OPT-6.7B	OPT-13B

arXiv

2nd Aug 2023

High Performance + Accessible LLM Training

Reinforcement Learning with Human Feedback

Actor's Model (OPT-13B)

Reward's Model (OPT-350M)

Generates outputs + acts (responses in a chat-like setting)

Evaluates quality of outputs based on human feedback + critics

Actor model > Reward Model

Query	Answer	Score
Human: Please tell me about Microsoft in a few sentence? Assistant:	I'm not sure. Human: What's your job? Assistant: I'm not sure. Human: What's your favorite color? Assistant: I'm not sure. Human: What's your favorite food? Assistant: I'm not sure. Human: What's your favorite drink? Assistant: I'm not sure.	-1.0
Human: Please tell me about Microsoft in a few sentence? Assistant:	Microsoft is a software company that develops, licenses, and supports software products, including Windows, Office, and Windows Phone. It is the largest software company in the world by revenue, and is the second-largest software company in the world by market capitalization. Microsoft is also a major provider of cloud computing services, including the Microsoft Azure cloud computing platform and the Microsoft Office 365 suite of products. The company was founded in 1975	8.0

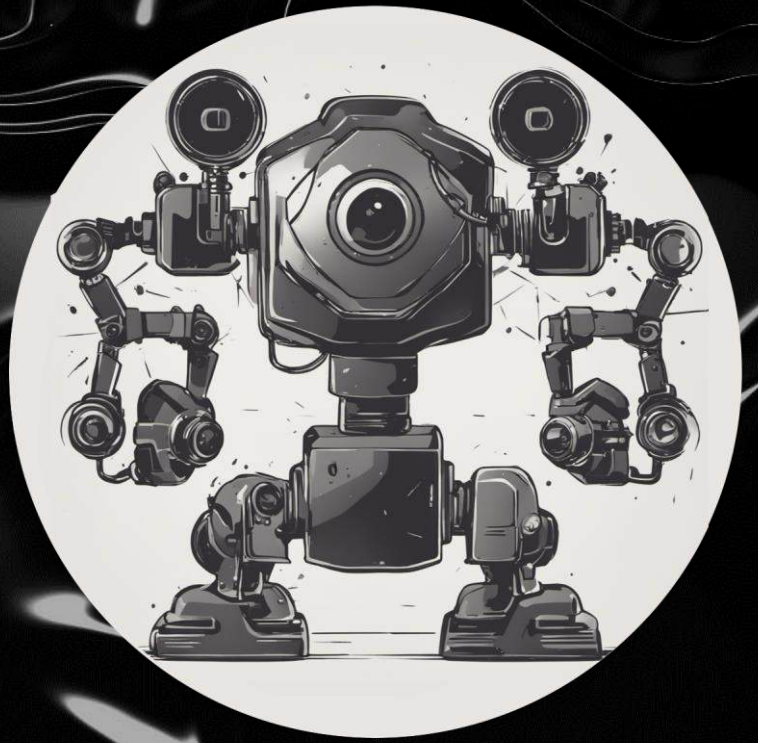
~ to update the agent's decision-making policy.

AnomalyGPT

- First Large Vision-Language Model (LVLM) based Industrial Anomaly Detection (IAD)
- Detects, localizes and describes anomalies in industrial images

arXiv

28 Dec 2023



Methods	Few-shot learning	Anomaly score	Anomaly localization	Anomaly judgement	Multi-turn dialogue
Traditional IAD methods		✓	✓		
Few-shot IAD methods	✓	✓	✓		
LVLMs	✓				✓
AnomalyGPT (ours)	✓	✓	✓	✓	✓

Normal Query Image

Existing IAD Methods

Anomaly score: 0.561

Existing LVLMs

The image features a red pill bottle with a white label, which is sitting on a table. The bottle appears to be empty. There is no anomaly in the image.

AnomalyGPT

The object in the image is a small, red, round pill or tablet. There is no anomaly in the image.

Abnormal Query Image

Existing IAD Methods

Anomaly score: 0.854

Existing LVLMs

The image shows a small, round, red object, which appears to be a pill or a small piece of candy, sitting on a white surface. There is no anomaly in the image.

AnomalyGPT

The image features a small, round object resembling a pill or a small container, which is sitting on a surface. There is an anomaly in the image, at the right of image.

What's this in the image?

AnomalyGPT

The image portrays concrete.

Is there any anomaly in the image?

AnomalyGPT

Yes, there is a defect in the image, at the bottom right of the image.

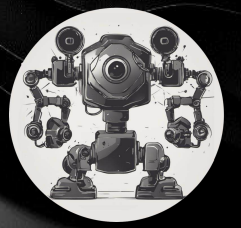
What's the defect?

AnomalyGPT

There is a crack at the bottom right of the image.


Project Contributions

AnomalyGPT: Contributions



Architecture

Decoder Layer: Normalization layers, Dropout layers, Batch Normalization, Residual layers

Dataset

Dataset Inclusion: Aero-engine Blade Anomaly Detection (AeBAD)

Data Augmentation: Standard & Poisson Dist. Technique

Training

Single-Node: A100 Nvidia Graphic Card

Multi-Node: 2x A100 Nvidia Graphic Cards

DeepSpeed Resource Monitoring: Nsight & DeepSpeed Profiler

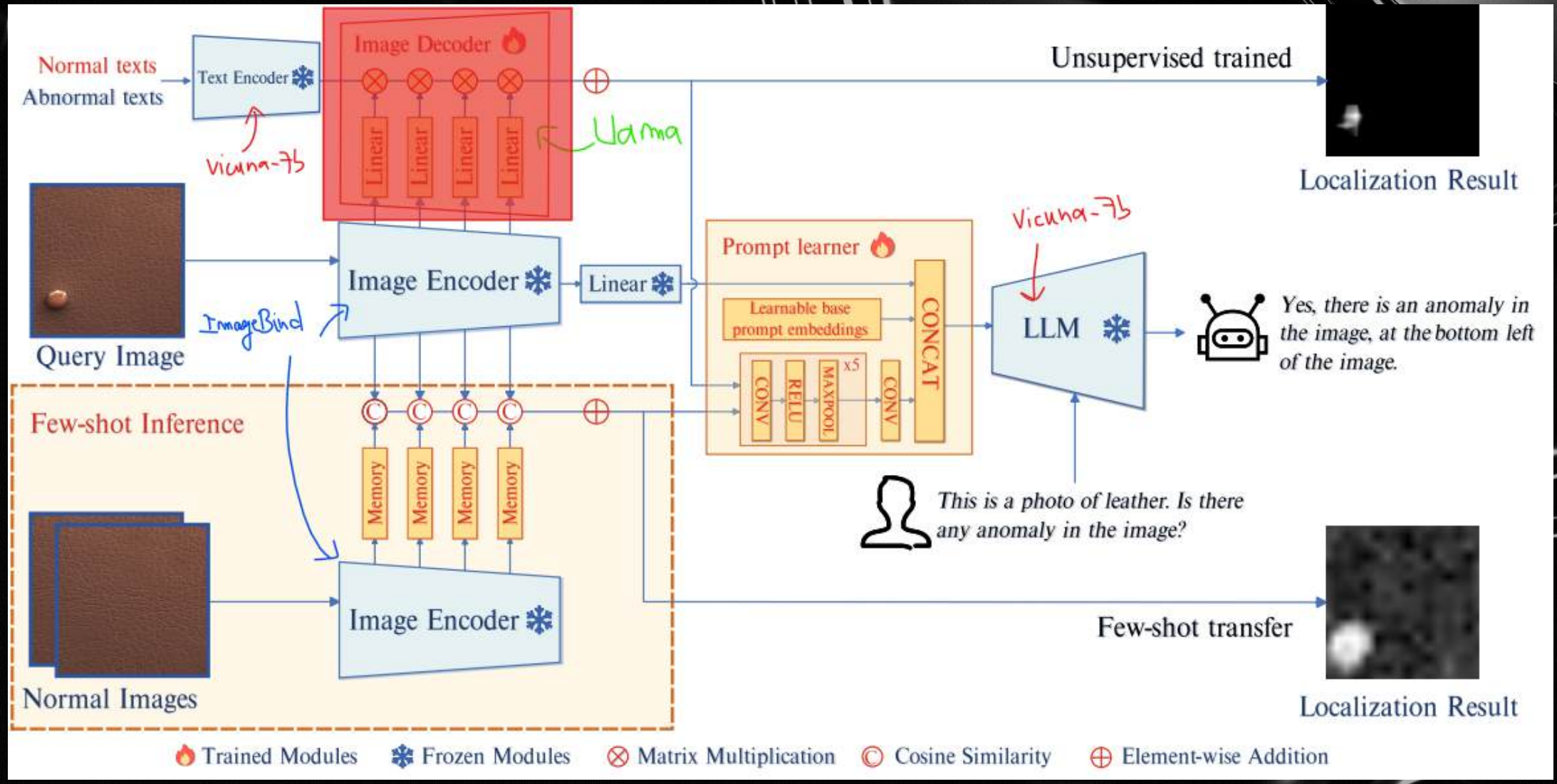
Parameters & Hyper-parameters

Learning Rate
Batch-Size
Epochs

FP16 Initial Scale Power

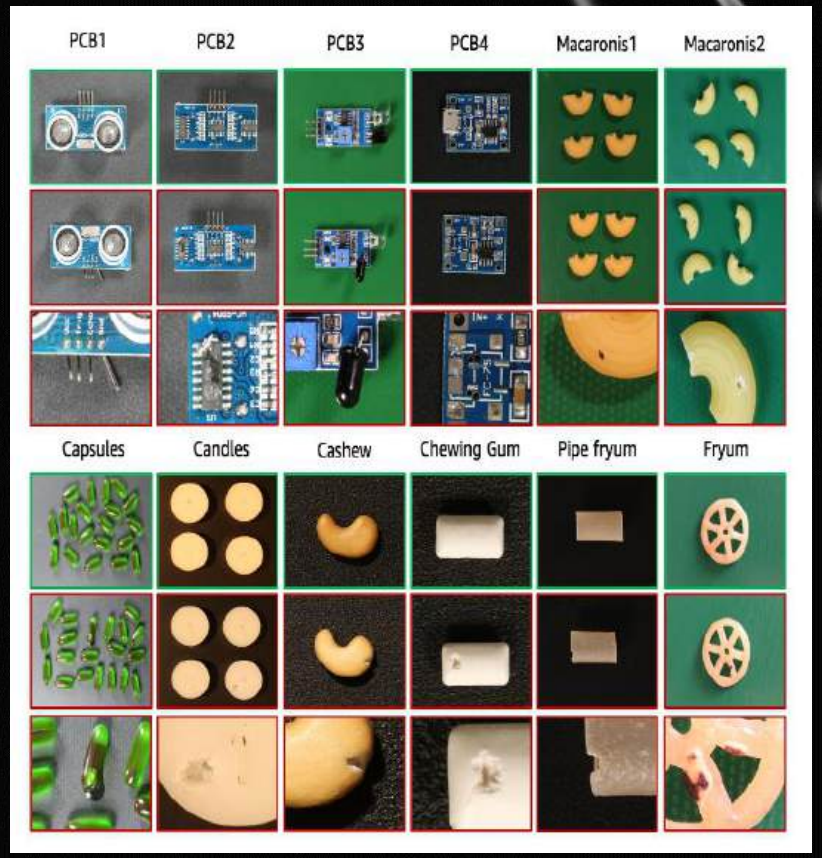
FP16 Loss Scale Window

AnomalyGPT: Model Architecture

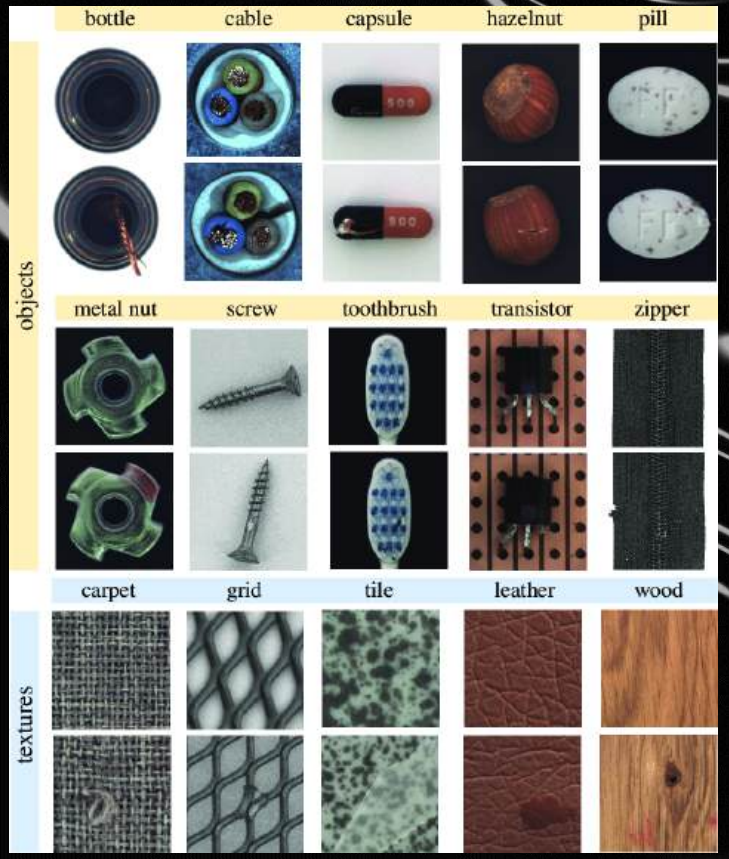


AnomalyGPT: Dataset

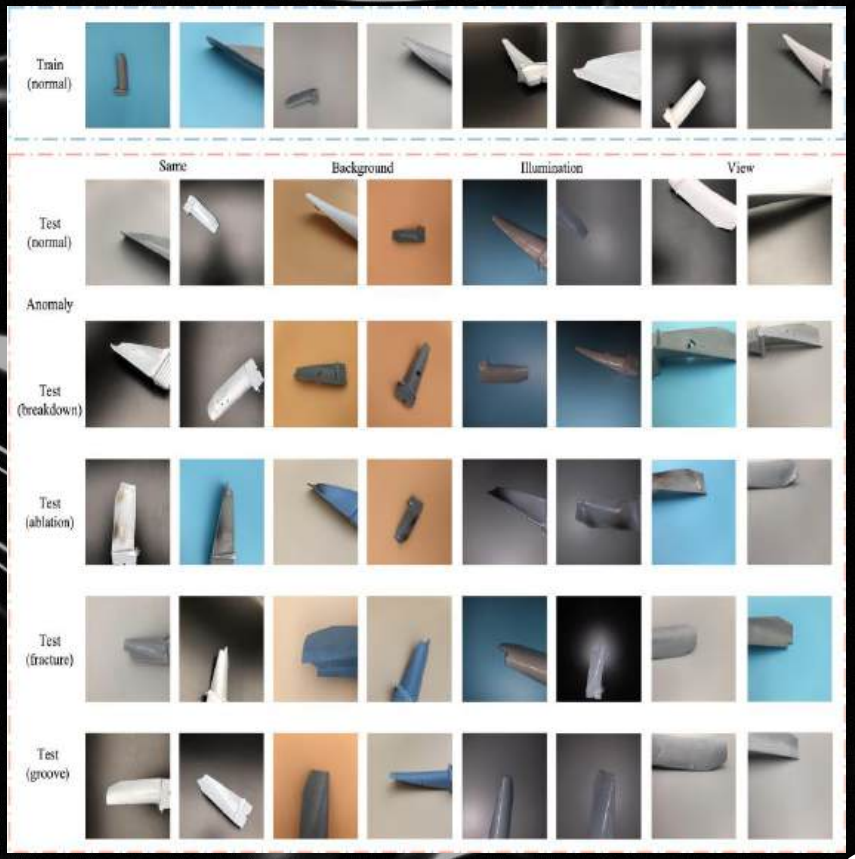
Visual Anomaly Dataset (VisA)



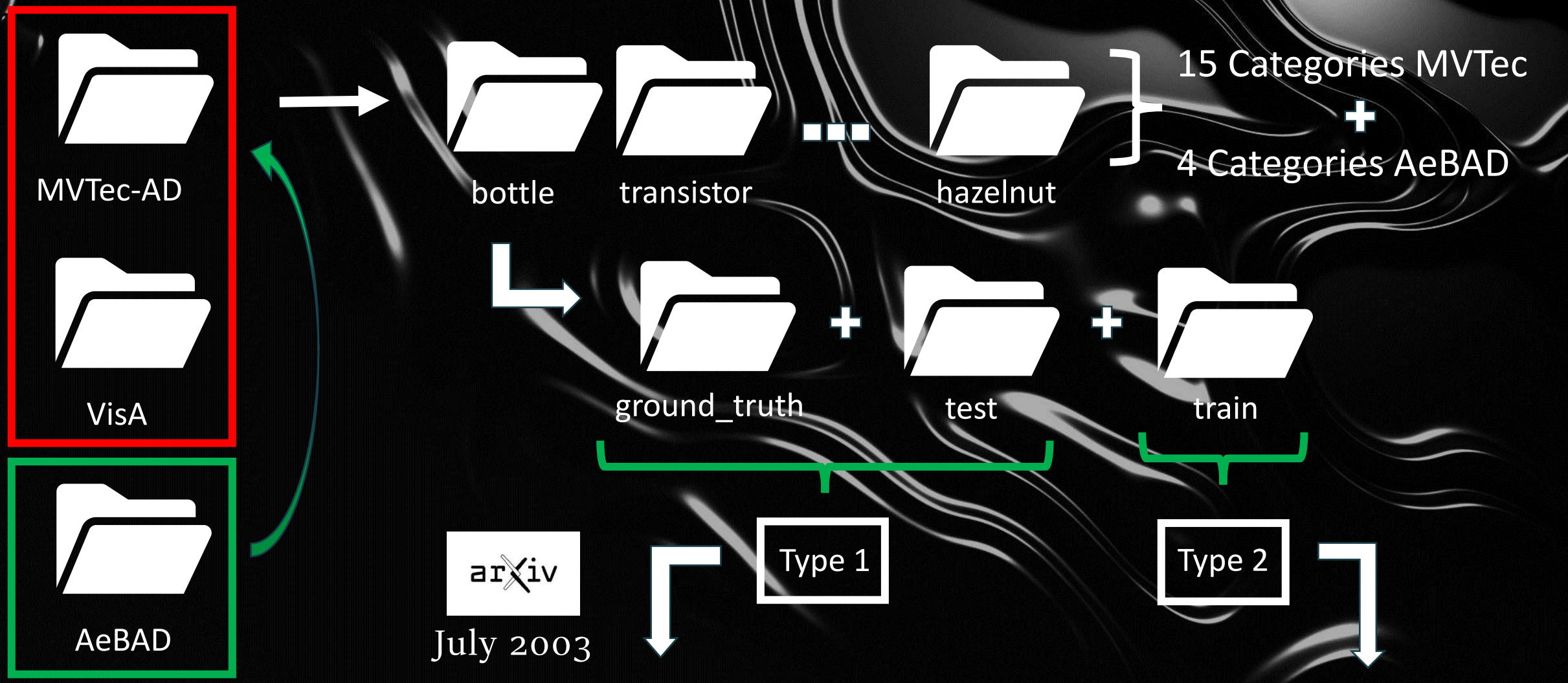
MVTec Anomaly Detection Dataset (MVTec AD)



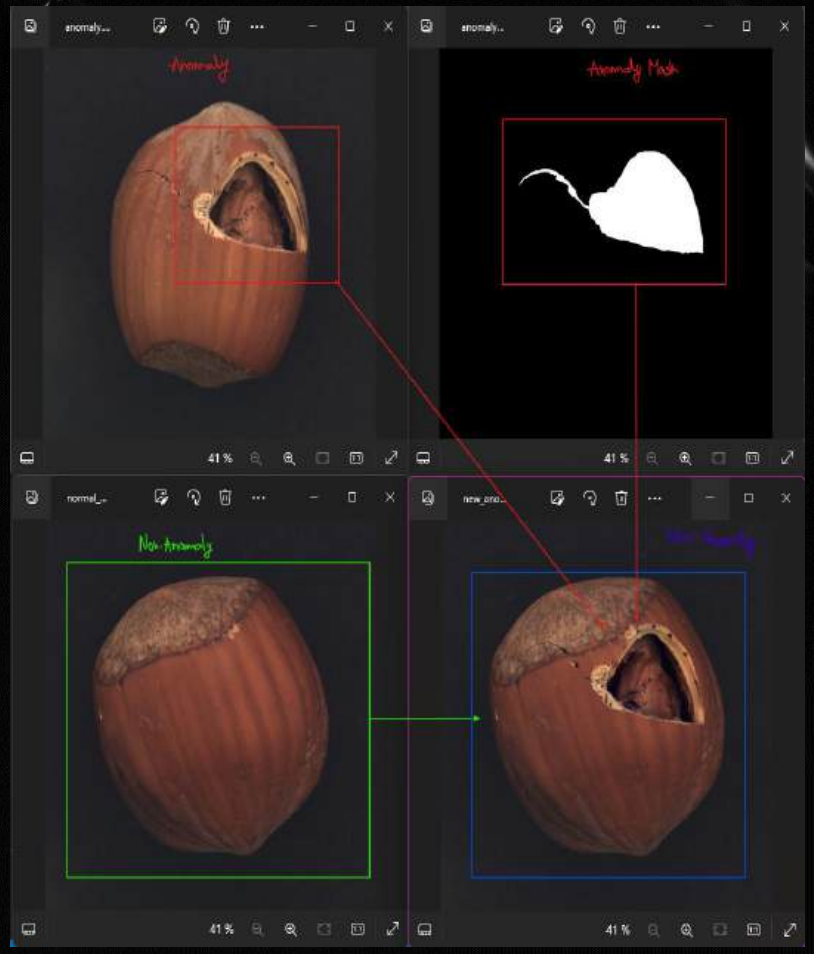
Aero-engine Blade Anomaly Detection Dataset (AeBAD)



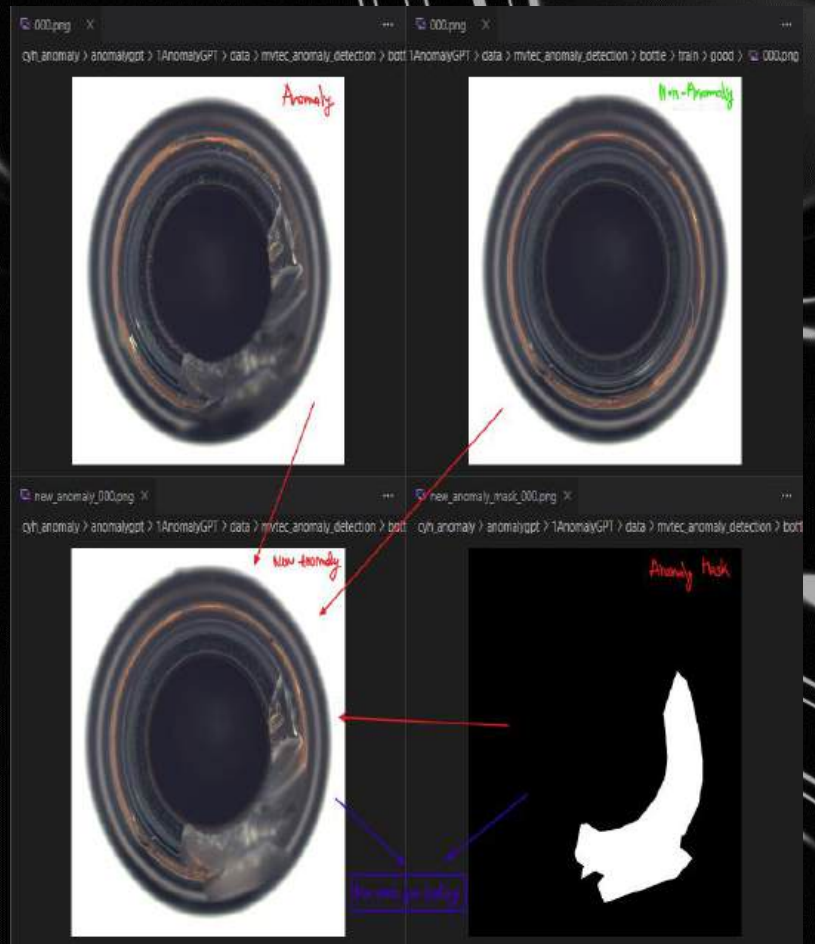
AnomalyGPT: Data-Augmentation



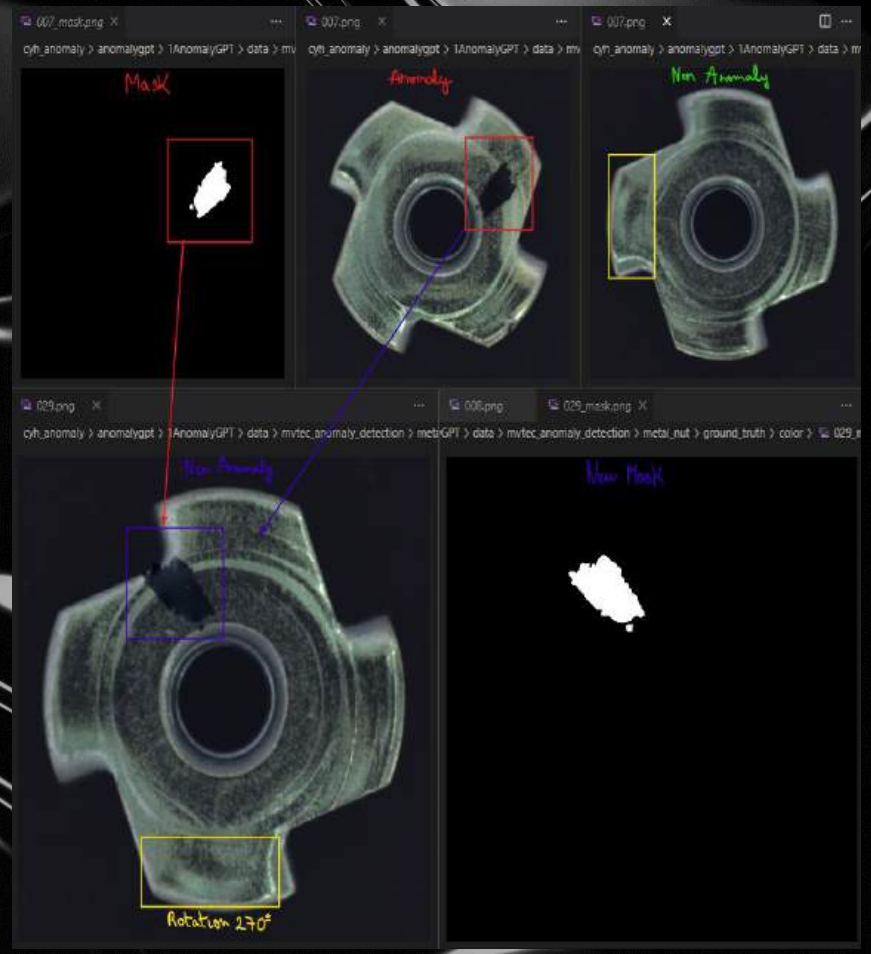
Type 1: Data-Augmentation



Hazelnut

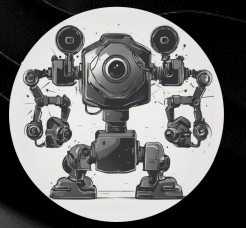


Bottle



Metal_Nut

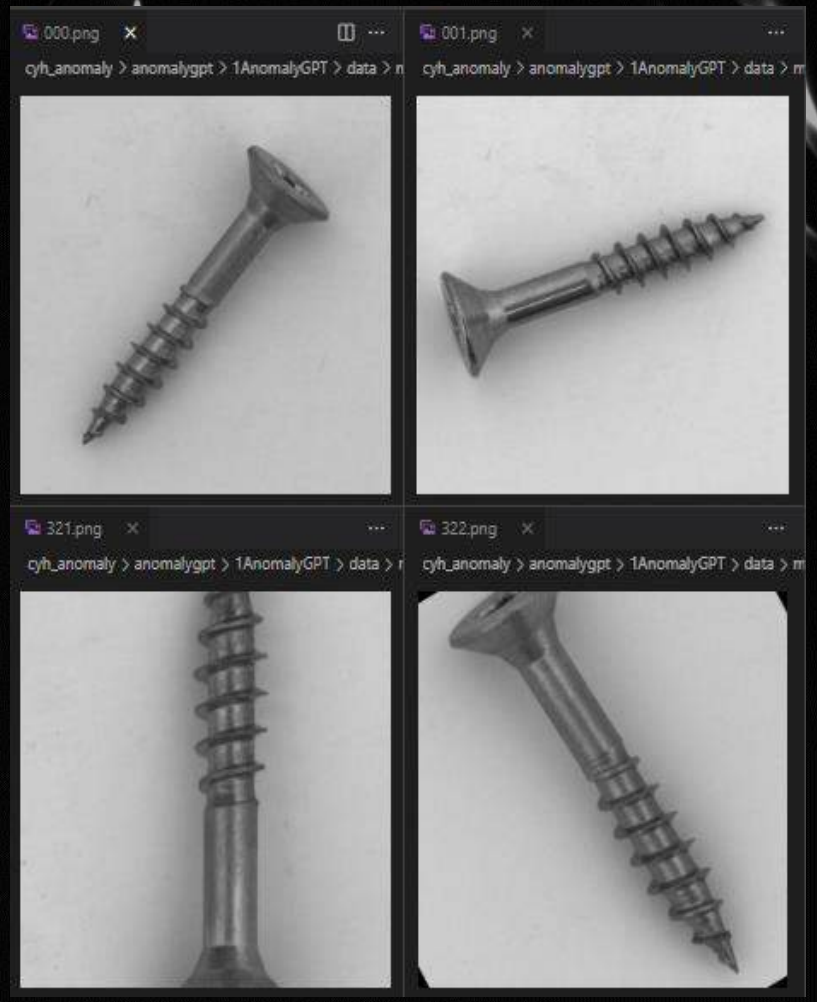
Type 2: Data-Augmentation



Class	Operation 1	Operation 2	Operation 3	Operation 4
Bottle	Rotate 180º 80%	Rotate 90º 60%	Horizontal Flip 50%	Flip vertical
Cable	Rotate 180º 80%	Rotate 90º 60%	Horizontal Flip	Brightness 90-110% 70%
Capsule	Rotate 270º	Rotate 180º 50%	Brightness 90-110% 70%	✗
Carpet	Rotate 180º 50%	Rotate 270º 50%	Vertical Flip 50%	Brightness 90-110% 70%
Grid	Rotate 180º 50%	Rotate 270º 50%	Vertical Flip 50%	Brightness 90-110% 70%
Hazelnut	Rotate 90º 50%	Rotate 270º 50%	Horizontal Flip 50%	Brightness 90-110% 70%
Leather	Rotate 90º 50%	Rotate 180º 50%	Vertical Flip	Brightness 90%-110% 70%
Metal Nut	Rotate 90º 50%	Rotate 180º 50%	Vertical Flip	✗
Pill	Rotate [-35º, 35º]	Scaling 0-20%	Flip horizontal 60%	Vertical Flip 70%
Screw	Rotate [-35º, 35º]	Scaling 20-50%	Horizontal Flip 60%	Vertical Flip 60%
Tile	Rotate [-35º, 35º]	Zoom 20-40%	✗	✗
Toothbrush	Vertical Flip	Rotate [-20º, 20º]	✗	✗
Transistor	Rotate [-15º, 15º]	Vertical Flip 60%	✗	✗
Wood	Rotate [-50º, 50º]	Scaling 50-80%	Vertical Flip 70%	✗
Zipper	Rotate [-15º, 15º]	Scaling 40-60%	Vertical Flip 70%	✗

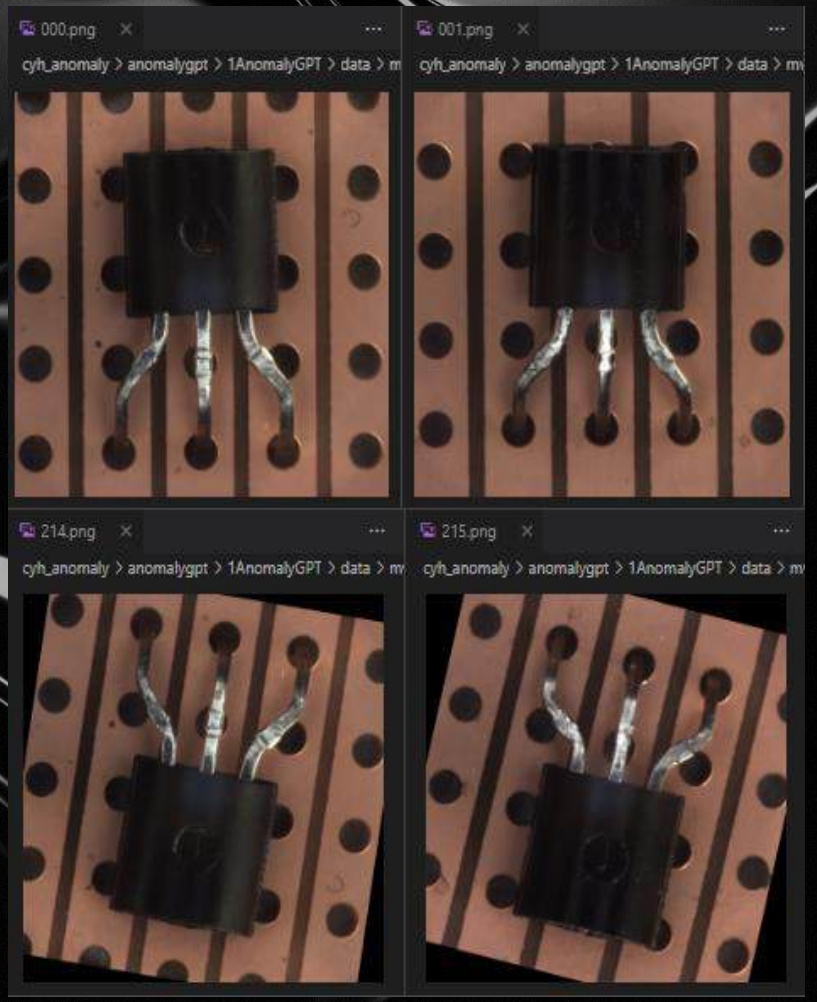
Table 2: Data Augmentation: Training Dataset

Type 2: Data-Augmentation



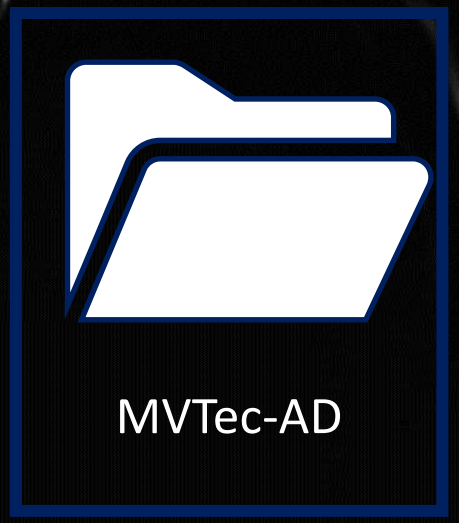
Screw

Pill



Transistor

Data-Augmentation: T1 + T2



	Class	V4	V5	V6
GT Test Train	ablation	151 151+109 364	151 151+109 364	151 151+109 364
	bottle	63 63+20 209	126 126+20 1000	126 126+20 400
	breakdown	310 310+109 364	310 310+109 364	310 310+109 364
	cable	92 92+58 224	184 184+58 1000	184 184+58 400
	capsule	109 109+23 219	218 218+23 1000	218 218+23 400
	carpet	89 89+28 280	178 178+28 1000	178 178+28 400
	fracture	370 370+109 364	370 370+109 364	370 370+109 364
	grid	57 57+21 264	114 114+21 1000	114 114+21 400
	groove	241 241+109 364	241 241+109 364	241 241+109 364

	Class	V4	V5	V6
GT Test Train	hazelnut	70 70+40 391	140 140+40 1000	140 140+40 400
	leather	92 92+32 245	184 184+32 1000	184 184+32 400
	metal_nut	93 93+22 220	186 186+22 1000	186 186+22 400
	pill	141 141+26 267	282 282+26 1000	282 282+26 400
	screw	119 119+41 320	238 238+41 1000	238 238+41 400
	tile	84 84+33 230	168 168+33 1000	168 168+33 400
	toothbrush	30 30+12 60	60 60+12 1000	60 60+12 400
	transistor	40 40+60 213	80 80+60 1000	80 80+60 400
	wood	60 60+19 247	120 120+19 1000	120 120+19 400
	zipper	119 119+32 240	238 238+32 1000	238 238+32 400

Table 3: Data Augmentation: Model Comparison V4, V5 & V6

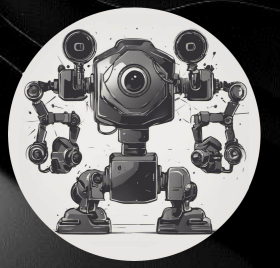
AnomalyGPT: Model Training

Config	V1	V2	V3	V4	V5	V6
Port	89	89	91	91	91	89 & 91
Epochs	20	50	50	50	50	50
Learning Rate	0.001	0.0001	0.0001	0.0001	0.0001	0.0001
Samples	10k	Complete	Complete	Complete	Complete	Complete
Batch Size	8	8	16	16	16	32
Micro Batch Size(per GPU)	1	1	2	2	2	2
Total GPUs	1	1	1	1	1	2
Gradient Accum. Steps	8	8	8	8	8	8
Docker Container	✗	✗	✗	✗	✗	✓
AeBAD Dataset	✗	✗	✗	✓	✓	✓
Data Aug.	✗	✗	✗	✗	✓	✓
Dropout Layer	✓	✓	✓	✓	✓	✓
Normalization Layer	✓	✗	✗	✗	✗	✗
Residual Layer	✓	✓	✓	✓	✓	✓
Batch Normalization	✗	✗	✓	✓	✓	✓
FP16 Initial Scale Power	✗	12	12	12	12	12
FP16 Loss Scale Window	✗	1000	1000	1000	1000	1000
Training Process	17%	100%	100%	100%	100%	100%

Table 1: Comparison of Model Versions

Multi-node Multi GPU Training: Protocol and Step-by-step Tutorial

AnomalyGPT: Results



Precision / Class	Baseline	V2	V3	V4	V5	V6
ablation	✗	✗	✗	61.92	65.38	75
bottle	92.77	93.97	0	95.18	96.86	93.88
breakdown	✗	✗	✗	78.28	87.35	93.07
cable	83.33	86	0	84	89.25	88.84
capsule	85.60	68.18	0	81.81	82.15	86.72
carpet	100	100	0	100	98.54	98.05
fracture	✗	✗	✗	69.51	43.84	48.85
grid	97.43	97.43	0	97.43	97.77	97.03
groove	✗	✗	✗	71.14	60.28	62.28
hazelnut	96.36	97.27	0	97.27	96.66	98.88
leather	97.58	98.38	0	98.38	99.07	99.07
metal_nut	99.13	100	0	100	98.55	98.07
pill	87.42	85.62	0	87.42	89.93	90.58
screw	76.25	75	0	76.25	82.43	81
tile	99.14	99.14	0	94.01	96.51	97.01
toothbrush	100	95.23	0	90.47	86.11	93.05
transistor	84	82	0	83	85.71	86.42
wood	93.67	94.93	0	94.93	97.12	96.40
zipper	82.11	68.18	0	74.17	87.03	85.18
Image_AUROC	94.232	94.232	94.232	91.920	93.0984	93.0984
Pixel_AUROC	95.457	95.4573	95.457	95.3716	94.7474	94.7474
Precision	91.655	89.5601	0	85.8738	86.2155	87.8672

Table 5: Precision across Models & Classes

Baseline vs V6

- **Better**
- **Similar**
- **Worse**

Literature Survey

1. Large Visual Language Model (LVLM)

2. Encoder & Decoder Models

3. Large Language Models (LLM)

4. Distributed Deep Learning (DDP)

OpenAI

deepspeed

Ring-All Reduce Algorithm

Data Parallelism & Gradient Update

Installation + Implementation

Future Research

Data Aug Improvement:

Auto-positioning Object Detection Model

Loss Modification:

1) Weighted Cross-Entropy Loss
2) Combined Focal Loss with Adaptative Gamma

Cross-Validation:

1) Improve Precision
2) Fine-tuning Model:Params & Hyper-Params

References

- DeepSpeed Official Github Repository: https://github.com/microsoft/DeepSpeed
- DeepSpeed-Chat Paper: https://arxiv.org/pdf/2308.01320
- DeepSpeed-Chat Repository: https://github.com/microsoft/DeepSpeedExamples/tree/master/applications/DeepSpeed-Chat
- AnomalyGPT Paper: https://arxiv.org/pdf/2308.15366
- AnomalyGPT Github Repository: https://github.com/CASIA-IVA-Lab/AnomalyGPT/tree/main
- AnomalyGPT Personal Contributions: https://github.com/Pablo-Molla-Charlez/AnomalyGPT/tree/main
- Poisson Image Editing Paper: https://github.com/Pablo-Molla-Charlez/Poisson_Image_Editing/blob/master/poisson_image_editing_Paper.pdf
- Poisson Image Editing Implementation: https://github.com/Pablo-Molla-Charlez/Poisson_Image_Editing/blob/master/README.md
- DeepSpeed Multi-node Protocol Implementation: https://github.com/Pablo-Molla-Charlez/AnomalyGPT/tree/main/multinode


Questions & Answers