INSTITUT POLYTECHNIQUE DE PARIS

Final Report

Out-of-Distribution Detection

Jiwon KANG
Mochamad Ardiansyah NUGRAHA

SIC7002 Artificial Intelligence for Data Science
M1 Electrical Engineering for Communications and Information Processing - Datapac
Institut Polytechnique de Paris
21 June 2024

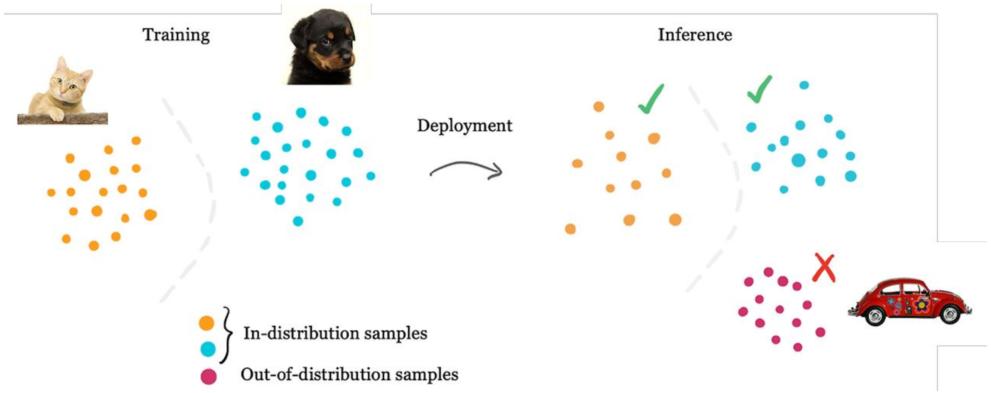
Introduction

What is OOD (Out-Of-Distribution) detection?

Determine whether a given new data point matches the distribution of the existing ID (In-Distribution) data or not.

Why is it important?

Ensures machine learning models remain reliable and safe when encountering unexpected data.



State-of-the-Art – Generalized OOD Detection

Background

- OOD is critical to ensure the reliability and safety of ML systems i.e. autonomous driving
- Isolated development of similar problems, such as anomaly detection (AD), novelty detection (ND), open set recognition (OSR), outlier detection (OD), are leading to confusion
- Incomplete summarization of OOD detection

Objective

- To present a unified framework called **generalized OOD detection,** each problem (AD, ND, OSR, OD, OOD) is considered as subtask
- To provide comprehensive discussion of methods from other subtasks
- To identify open challenges and potential research directions

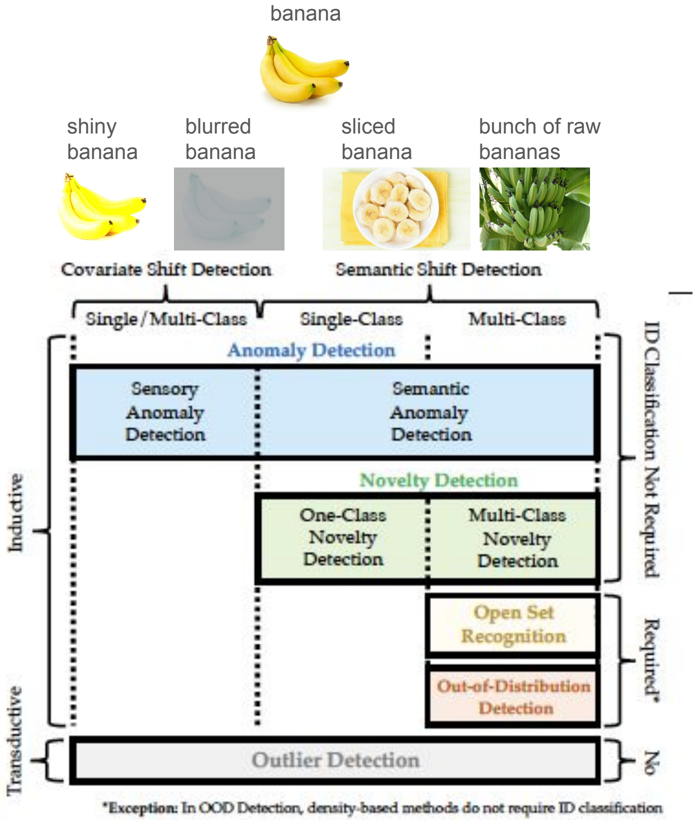
State-of-the-Art – Generalized OOD Detection

Baseline of OOD detection taxonomy

1. Distribution shift
    a. Covariate shift: **change in input** space while **label** space remains **constant** i.e. adversarial, style changes
    b. Semantic shift: introduction of **new categories** or **alteration** of existing ones, directly **impacting label space** and consequently the input space
2. ID classes: single or multiple
3. Is required ID classification?
4. Transductive or inductive task

ND is often **interchangeable with AD**, but ND is more concerned on **semantic** anomalies.

OOD detection is generally **interchangeable with OSR**.

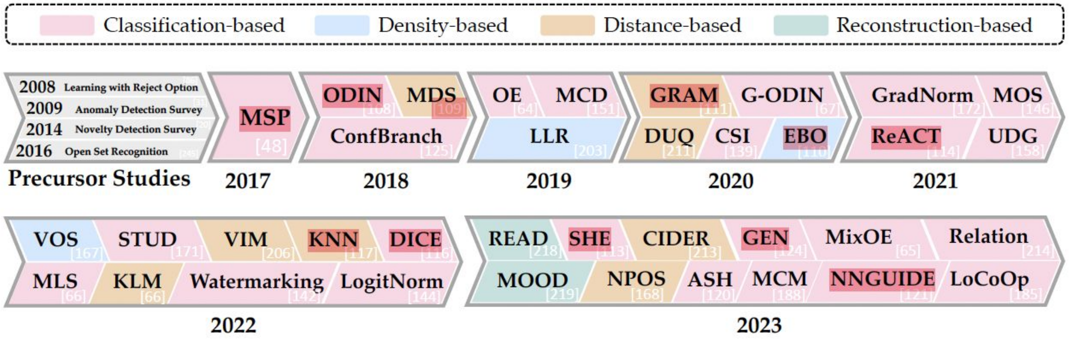


State-of-the-Art

Summary of all latest OOD detection methods

Table 1 Paper list for out-of-distribution detection.

Sections		
§ 3.1 Classification	§ 3.1.1 Output-based Methods	a: Training-free
		b: Training-based
	§ 3.1.1 Outlier Exposure	a: Real Outliers
		b: Data Generation
	§ 3.1.3: Gradient-based Methods	
	§ 3.1.4: Bayesian Models	
	§ 3.1.5: OOD for Foundation Models	
§ 3.2: Density-based Methods		
§ 3.3: Distance-based Methods		
§ 3.4: Reconstruction-based Methods		
§ 3.5: Theoretical Analysis		



Methodology

Scope

- Multi-class
- Semantic shift
- Near-OOD
- Post-hoc (Training-free)

Post-processor

No hyperparameter tuning

MSP - *Maximum Softmax Probabilities*: takes the maximum softmax probability over ID categories as the score

TempScale - *Temperature Scaling*: calibrates softmax probabilities with temperature scaling

MLS - *Maximum Logit Scaling*: uses the maximum logit

With hyperparameter tuning

EBO - *Energy-based OOD*: applies energy function to the logits to compute OOD score

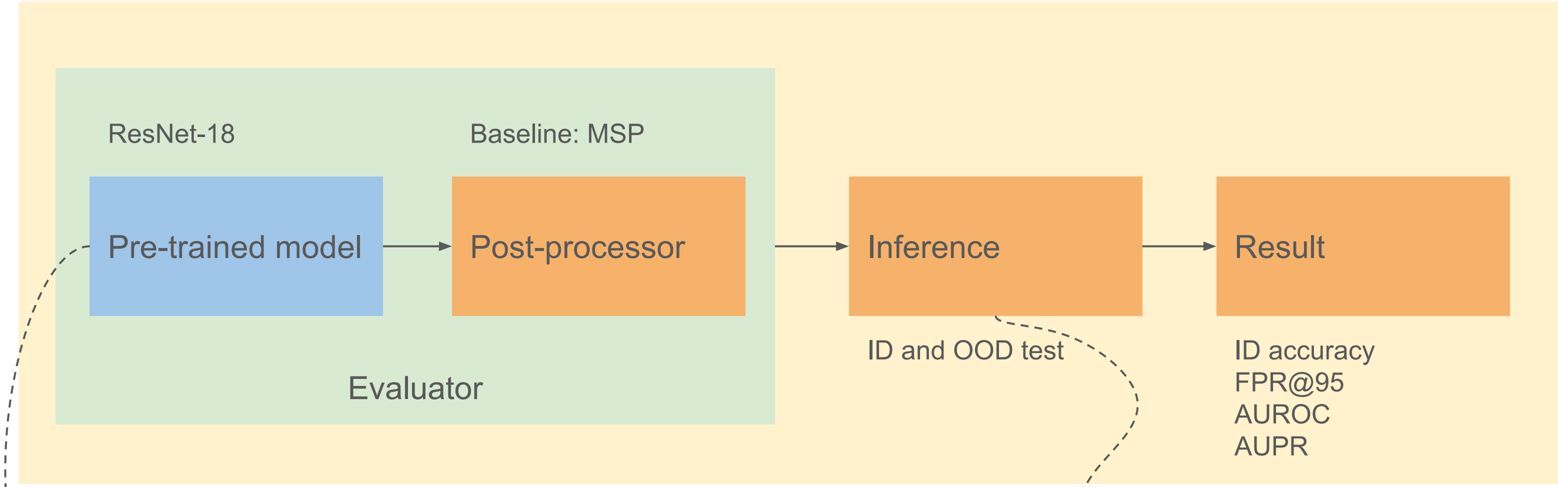
ODIN - *OOD Images in Neural Net*: introduces input preprocessing on top of TempScale

ReAct - *Rectified Activation*: rectifies feature vectors by thresholding their elements with a certain magnitude

Methodology

Obtain pre-trained models → Utilize OpenOOD library, set MSP as baseline → Experiment with other pre-processors → Compare performance

Methodology - Building Blocks



ResNet18
Epoch : 100
Loss function : Cross-entropy
Optimizer : Stochastic gradient descent
Momentum : 0.9
Learning rate : 0.1, with cosine annealing decay schedule
Batch size : 128

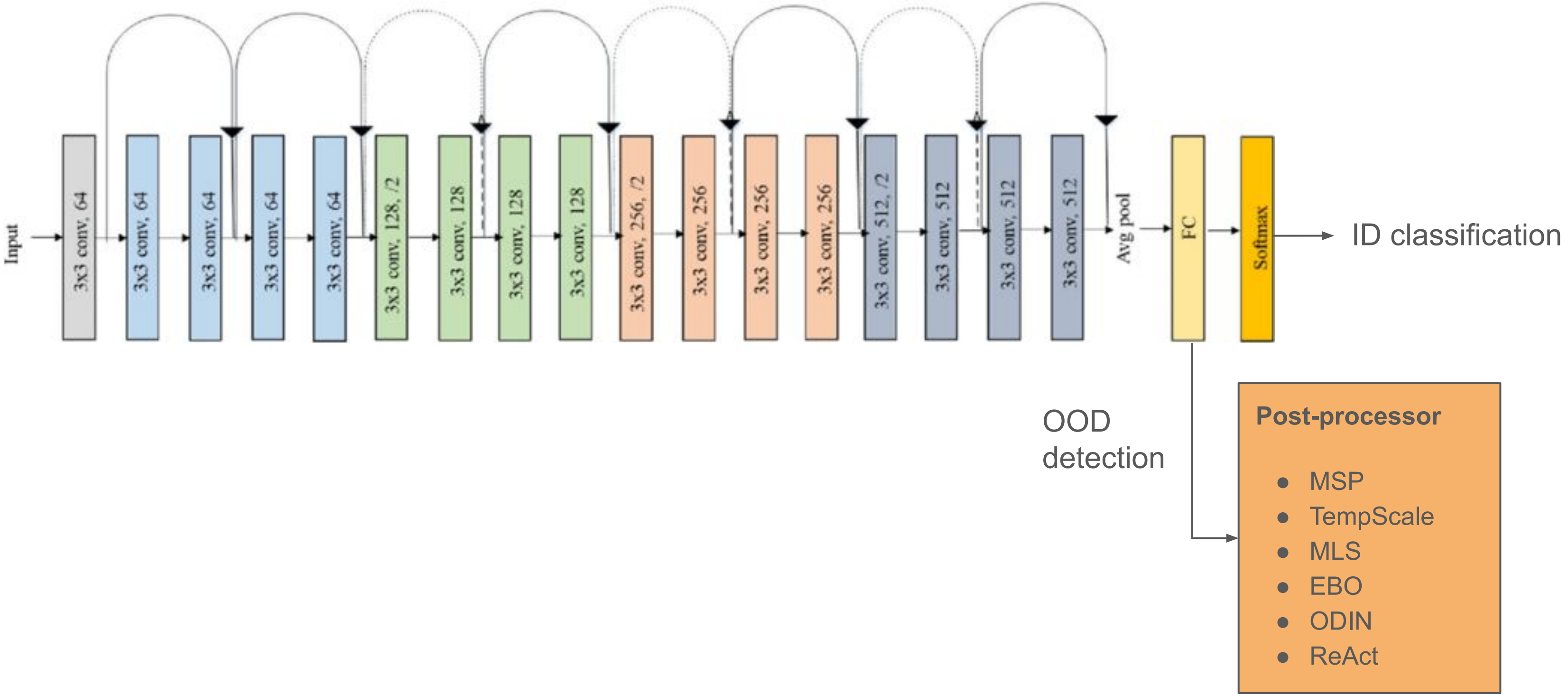
Datasets

1. ID: CIFAR-100
   OOD: CIFAR-10, TinyImageNet
2. ID: CIFAR-10
   OOD: CIFAR-100, TinyImageNet

Tools

- OpenOOD library
  https://github.com/Jingkang50/OpenOOD
- PyTorch

Methodology – ResNet-18 Architecture + Post-hoc OOD



Result – CIFAR-100

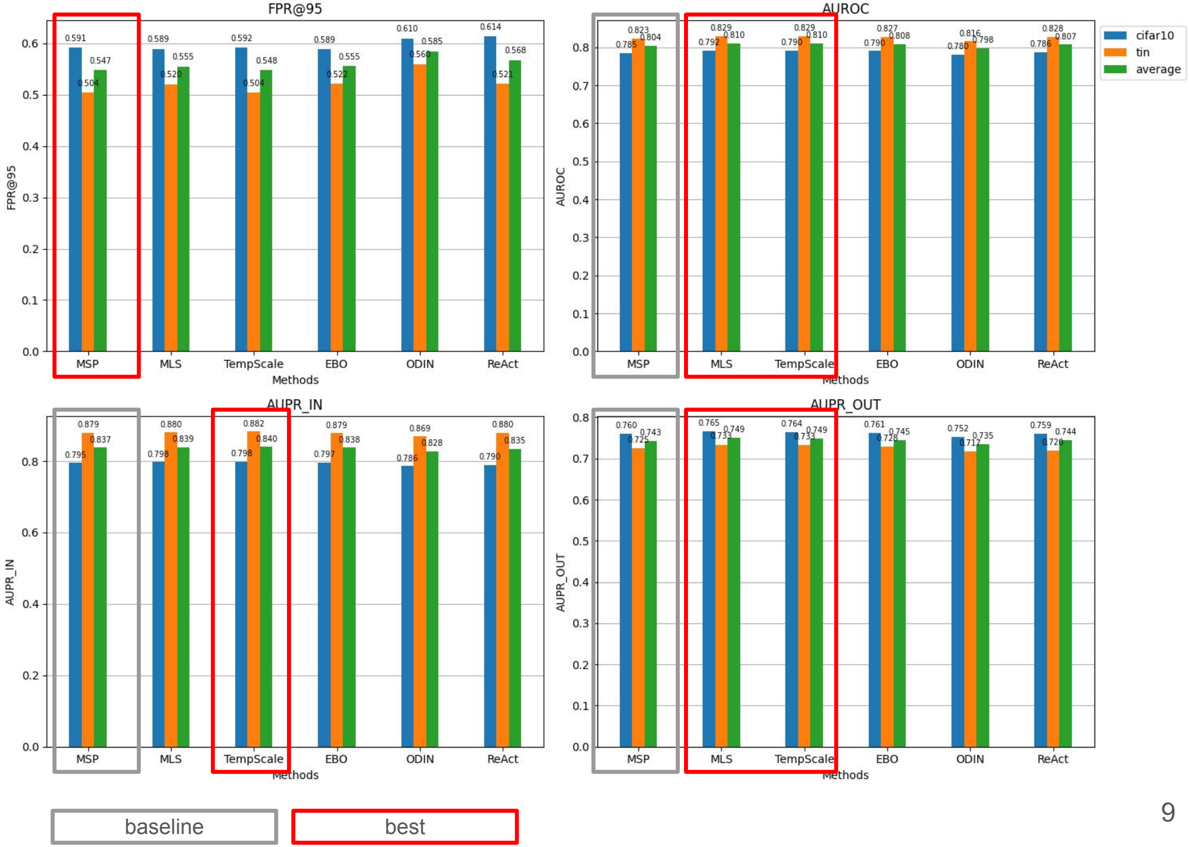
CIFAR-100 ID accuracy: 77.17%

OOD datasets: CIFAR-10, TinyImageNet

Average OOD scores vs baseline (MSP):

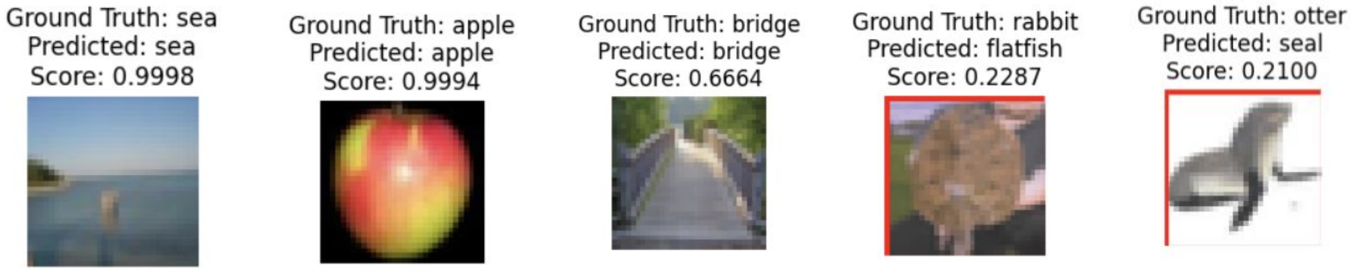
	Baseline (MSP) vs Average OOD scores	
Lowest FPR	0.54	
Highest AUROC	0.80	0.81 (MLS, TempScale)
Highest AUPR_IN	0.83	0.84 (TempScale)
Highest AUPR_OUT	0.74	0.75 (MLS, TemScale)

Since the **AUROC** and **AUPR** values among all methods share **very small differences**, we consider the **FPR as the primary score**.



Result – CIFAR-100 with MSP

ID CIFAR-100



OOD CIFAR-10

Result – CIFAR-10

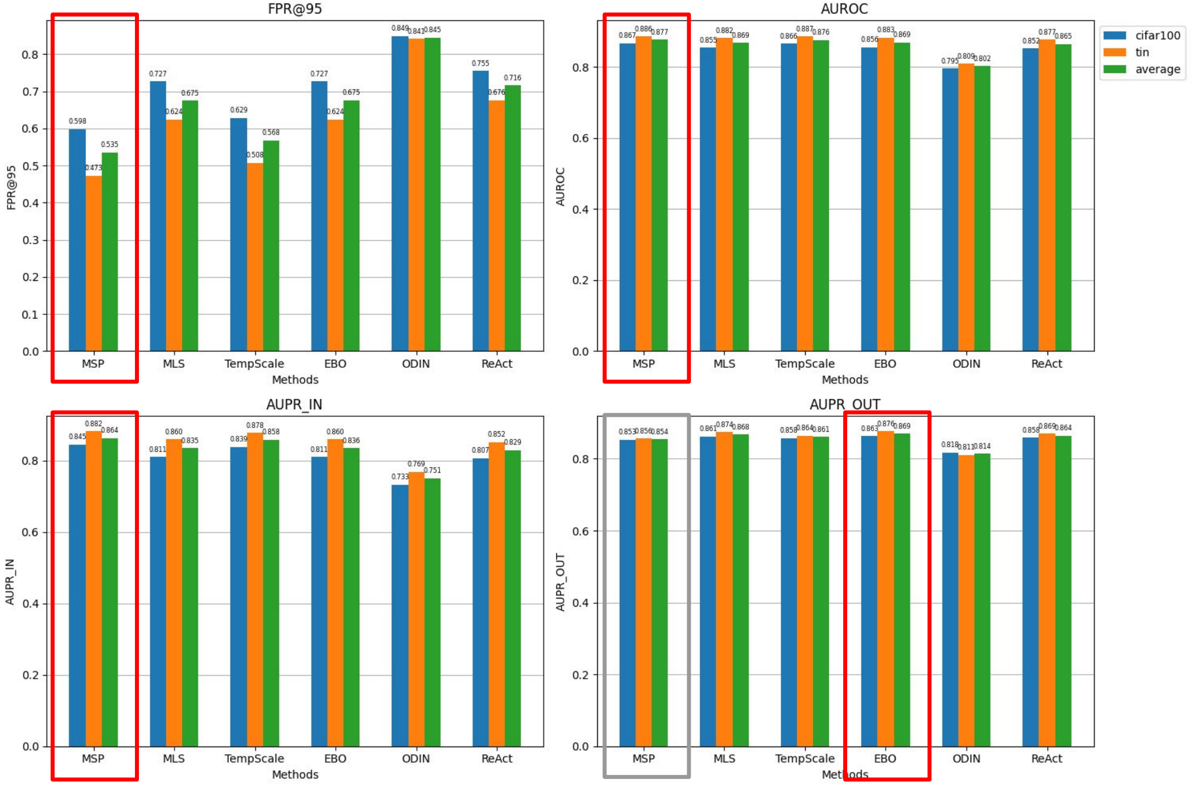
CIFAR-10 ID accuracy: 95.22%

OOD datasets: CIFAR-100, TinyImageNet

Average OOD scores vs baseline (MSP):

- **Lowest FPR**: 0.535 (MSP)
- **Highest AUROC**: 0.877 (MSP)
- **Highest AUPR_IN**: 0.864 (MSP)
- **Highest AUPR_OUT**: 0.869 (EBO) vs. 0.854 (MSP)

Overall, as baseline, **MSP performs better** than other post-processors.



baseline best

INSTITUT POLYTECHNIQUE DE PARIS

Conclusion

- ID **accuracy of CIFAR-10** is way better than CIFAR-100, **95.22%** vs. 77.17%.
- CIFAR-100: Temperature Scaling (TempScale) and Maximum Logit Scaling (MLS) share the **highest values** in terms of **AUROC** and **AUPR**. But, **the differences** with other pre-processors are also **very small**.
- CIFAR-10: Maximum Softmax Probability (MSP) **outperforms** other pre-processors.
- Dataset with **smaller number of ID class** seems can be equipped with **simpler post-processor** (i.e. CIFAR-10 with MSP).

Comments & Suggestions

- Due to time and computing **resource limitation**, this research only covered post-hoc methods.
- For real-life application, post-hoc-only methods **might not be sufficient**. It needs broader benchmark and data.
- Some OOD methods might be well-fit and tailored for **specific problems** (i.e. self-driving, cancer detection, etc). Hence, it is encouraged to explore other methods than post-hoc for future works.

INSTITUT POLYTECHNIQUE DE PARIS

Reference

1. Jingkang Yang et al,, “Generalized Out-of-Distribution Detection”, arXiv:2110.11334v3, 2024.
2. Jingyang Zhang et al., “OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection”, arXiv:2306.09301v2, 2023.
3. Jingkang Yang et al., “OpenOOD: Benchmark Generalized Out-of-Distribution Detection”, arXiv:2210.07242v1, 2022.

Questions

1. **What are the advantages and limitations of using MSP as a baseline method for OOD detection?**
2. **Why do models perform better on CIFAR-10 compared to CIFAR-100 in OOD detection?**
3. **What are the main issues commonly encountered in OOD detection?**
4. **Referring to generalized OOD detection framework, which OOD detection subtask that is suitable for product quality control in manufacture? Why?**