DNN-GP: Diagnosing and Mitigating Model's Faults Using Latent Concepts

Shuo Wang
Shanghai Jiao Tong University

Hongsheng Hu
CSIRO's Data61

Jiamin Chang
University of New South Wales
CSIRO's Data61

Benjamin Zi Hao Zhao
Macquarie University

Qi Alfred Chen
University of California, Irvine

Minhui Xue
CSIRO's Data61

Abstract

Despite the impressive capabilities of Deep Neural Networks (DNN), these systems remain fault-prone due to unresolved issues of robustness to perturbations and concept drift. Existing approaches to interpreting faults often provide only low-level abstractions, while struggling to extract meaningful concepts to understand the root cause. Furthermore, these prior methods lack integration and generalization across multiple types of faults. To address these limitations, we present a fault diagnosis tool (akin to a General Practitioner) DNN-GP, an integrated interpreter designed to diagnose various types of model faults through the interpretation of latent concepts. DNN-GP incorporates probing samples derived from adversarial attacks, semantic attacks, and samples exhibiting drifting issues to provide a comprehensible interpretation of a model's erroneous decisions. Armed with an awareness of the faults, DNN-GP derives countermeasures from the concept space to bolster the model's resilience. DNN-GP is trained once on a dataset and can be transferred to provide versatile, unsupervised diagnoses for other models, and is sufficiently general to effectively mitigate unseen attacks. DNN-GP is evaluated on three real-world datasets covering both attack and drift scenarios to demonstrate state-to-the-art detection accuracy (near 100%) with low false positive rates ($< 5\%$).

1 Introduction

Deep Neural Networks (DNN) have become integral to pushing the envelope of decision making performance in an increasing number [16] like health [8, 21], transportation [32, 44], and security-sensitive areas [46]. However, they have been demonstrated to make blunders, often due to issues related to Adversarial Robustness [12, 13, 33] and Concept Drift [18, 35]. A persisting challenge mitigating the effect of these blunders is the ability to interpret why these fault decisions were made and identify the the root cause [51].

Research Gaps. Despite the significant strides made in understanding the DNN-related errors, two salient gaps persist, the low-level abstraction of interpreting faults, and the lack of an integrated approach to resolving different faults.

Low-level Abstraction of Interpretations. Current techniques, such as Grad-CAM [15, 50], visualize the importance of image features to decision making with heat maps at the individual pixel-level per instance. While detailed, this approach does not reveal the high-level *root causes* of the misclassification. Effective interpretations necessitate operating at a higher level of abstraction, for example at a conceptual level [3, 7, 28, 36]. Unfortunately, these conceptual approaches have been developed to operate on tabular attribute data, posing significant challenges in extracting meaningful concepts from high-dimensional data, like images. An alternative approach involves using autoencoders [54] to transform high-dimensional input into an interpretable latent space. However, such free-form one-to-one mappings are hard to reliably capture consistent, recurring, and global features in the data. Thus, it is difficult to construct a comprehensive concept space.

Lack of Defense Integration. Existing solutions are always narrow-focused approaches to address specialized vulnerabilities or achieve particular generalizations of a given attack. Due to the arms-race development of machine learning attacks and their respective defenses, these proposals have not sought the breadth to concurrently address an array of vulnerabilities. A quintessential solution should boast the versatility to diagnose a myriad of models across diverse faults, solely equipped with knowledge of the clean or original dataset. For instance, a diagnostic methodology crafted to tackle adversarial perturbations should be equally adept at addressing semantic transformations leading to model inaccuracies. Moreover, the crux of the diagnostic process lies not just in identifying faults but in using this insight as a springboard to bolster fault mitigation, ultimately fortifying model resilience.

Research Question. Given these gaps, research question is *"to propose an interpreter that can develop a conceptual space to diagnose model's faults caused by adversarial/semantic attacks, and data drift, then to guide efficient mitigation."*

Motivation. To this end, we propose DNN-GP, an integrated interpreter to diagnose various types of vulnerabilities and

faults using latent conceptual interpretation. The motivation behind DNN-GP is threefold:

Global Consistent Conceptual Space. Our strategy relies on mapping high-dimensional input to a low-dimensional latent conceptual space that is both consistent and informative. Here, a global concept codebook plays a pivotal role in preventing free-form latent representations derived from ordinary autoencoders. It helps to direct the encoder towards a set of fixed, globally shared concepts (one-to-standard). Every concept in the global codebook essentially serves as a prototype, encapsulating a cluster of similar features from the dataset. One strength of the concept codebook is its ability to self-learn from training data, eliminating the need for prior domain-specific knowledge. This provides a compact and summarized representation of the data while ensuring that the latent space is interpretable through concepts, with each codebook concept offering insights into underlying patterns of the dataset.

Comprehensive Diagnosis. By deploying the encoder alongside the global concept codebook, we can seamlessly diagnose samples. This is achieved by encoding each sample into its latent representation, followed by aligning this representation with the global concepts from the codebook. Consequently, both the aligned concept index and the associated alignment distance serve as diagnostic indicators, effectively quantifying how the sample deviates from the established norms in the original dataset based on the hypothesis testing. We then apply further diagnostic methodologies to dissect samples stemming from existing adversarial attacks, semantic attacks and data drifting issues, named probing samples. As shown in Figure 1, DNN-GP provides three-tiered diagnostics: *(i)* Concept Patterns of Usage and Alignment: DNN-GP's encoder maps the input into a 2D latent representation, such as the aligned concept index matrix coupled with the corresponding alignment distance matrix. This enables the identification of aberrant distance distribution patterns and the detection of unusual or susceptible concept usages (red concepts and bars of Figure 1). *(ii)* Spatial Patterns in Latent Space: The 2D latent representation can also reveal abstract spatial patterns. For example, by comparing the average alignment distance matrices of original and attack datasets, we can unearth the underlying causes of model misclassifications. Visualization of these patterns suggests that perturbations causing misclassifications in gender often resemble makeup applications in facial areas. *(iii)* Pixel Pattern via Reconstruction: DNN-GP is adept at converting non-structural perturbations at the pixel level into structured patterns in the pixel space via reconstruction. These discernible patterns serve as a valuable reference to assist in understanding attacks.

Effective Mitigation. Utilizing the diagnostic insights from DNN-GP, we can devise conceptual strategies to bolster a model's resilience. For instance, feature selection can be applied to the latent conceptual representation to detect malicious or anomalous samples from benign ones. These subsequent mitigation strategies boast the benefits of being model- and attack-agnostic. As DNN-GP's countermeasures are rooted in the foundational latent conceptual representations derived from the original dataset, it sidesteps the vulnerabilities of current interpretation methods, which are susceptible to adversarial manipulations.

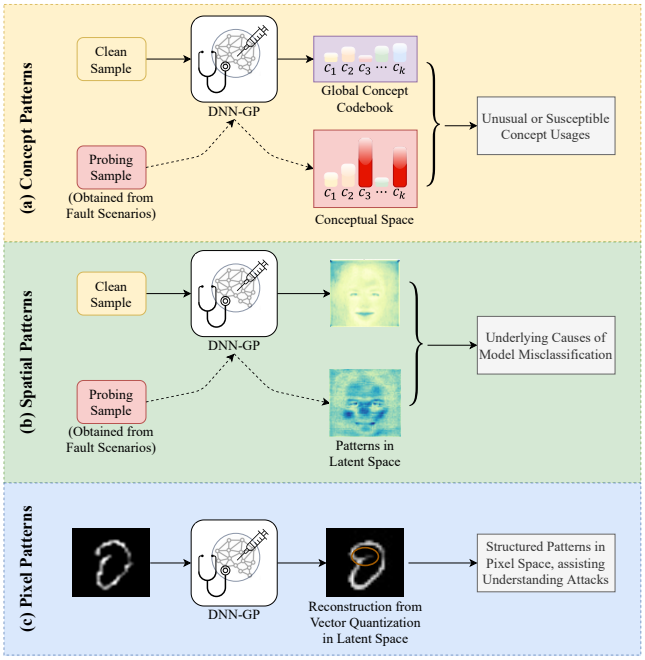


Figure 1: An illustration of DNN-GP's three-tiered approach to diagnose DNN faults and data quality issues.

Contributions. Key contributions are summarized as follows:
(i) Benchmark Dataset. We curate a benchmark dataset with a wide spectrum of model vulnerabilities, including adversarial attacks, semantic attacks, and data drift instances.
(ii) Integrated Diagnostic Tool. Introducing DNN-GP, a open diagnostic tool that elucidates DNN faults using patterns in the conceptual space. Once trained on a dataset, DNN-GP offers a continuous, versatile, and unsupervised diagnosis and mitigation across various models and faults.
(iii) Innovative Perspective. DNN-GP introduces a fresh lens to anchor probing samples within a consistent conceptual space, moving beyond instance-level model interpretations. This allows a handful of probing samples to catalyze insights into the underlying causes of model's faults to inform decisions on vulnerability evaluations, strengthen resilience, and/or formulate adversarial tactics.
(iv) Resilience Detection. We introduce an approach that harnesses latent concept representation to bolster model robustness through detection. This detection mechanism achieves state-of-the-art performance, nearing 100% with minimal false positives. It excels in an unsupervised, fault-agnostic, model-independent, and streamlined fashion. Notably, our method adeptly identifies even advanced attacks with minimal perturbations, outperforming existing solutions.

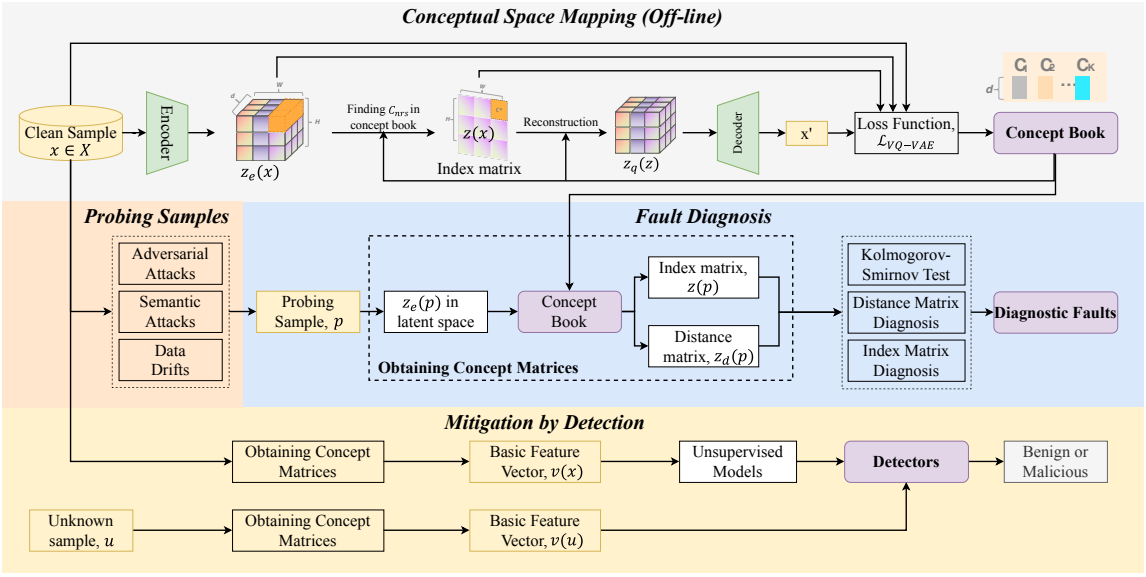


Figure 2: Functional overview of the DNN-GP. Observe how DNN-GP establishes a learned aligned concept book from a data distribution, leverages probing samples to diagnose a range of faults, and uses latent concept alignments to mitigate faults.

2 DNN-GP

Our system comprises four components in Figure 2.
***(i)* Probing Samples.** To unravel the intricacies of DNN behavior, we delve into three distinct fault scenarios from which we derive our 'probing samples'. The scenarios span Adversarial Examples, Semantic Attacks, and Data Drift.
***(ii)* Conceptual Space Mapping.** Leveraging Vector Quantization, we train an encoder and establish a global codebook from the original dataset. This framework allows us to craft latent conceptual representations for samples, achieved by aligning their encoded embeddings to global concept markers. This process is enriched with a clustering-driven alignment technique, ensuring a finer granularity of the conceptual space.
***(iii)* Fault Diagnosis.** Delving deeper into the outcomes from the second component, we employ statistical methods on the aligned concept index matrices and their respective distance matrices. This analytical approach aids in elucidating the DNN's response to the probing samples.
***(iv)* Mitigation.** Given diagnostic findings, we advocate for mitigation strategies, with an emphasis on identifying anomalous or malicious samples rooted in latent concept patterns.

2.1 Probing Samples

We use three fault settings to obtain "probing samples", including *adversarial attacks* for both white-box (e.g., Fast Gradient Sign Method (FGSM) [20], Projected Gradient Descent (PGD) [39], DeepFool [42], JSMA [45], Carlini & Wagner C&W [14] and Pixel Attack [29]) and black-box (e.g., Square Attack [5]), *semantic attacks* (e.g. contrast change [19], rotation [17], color-shift [25]) and *data drifts* (e.g., MNIST-C [43] and CIFAR-C [24]). By analyzing probing samples, we can gain a deeper understanding of patterns resulting in faults, and develop strategies to mitigate vulnerabilities. Detailed settings of probing samples are in Appendix A.

2.2 Conceptual Space Mapping

2.2.1 Construction of Concept Codebook

Vector Quantized Variational AutoEncoders (VQ-VAE) [47] are used to learn discrete conceptual representations of data and a global conceptual dictionary or codebook. In general, a VQ-VAE consists of three components, the encoder $\mathrm{En}(\cdot)$, decoder $\mathrm{De}(\cdot)$, and codebook $C = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_K\}$. The codebook C defines the commonly-shared latent embedding space $C \in \mathbb{R}^{K \times d}$, consisting of K categorical embedding items with d dimensions, which we call *concept* item, i.e., $\boldsymbol{c}_i \in \mathbb{R}^d$, $i \in \{1, 2, \ldots, K\}$. The encoder is a non-linear mapping from the input instance $\boldsymbol{x}$ in the pixel space to the latent representation $\boldsymbol{z}_e(\boldsymbol{x}) \in \mathbb{R}^{W \times H \times d}$. Specifically, $W \times H$ latent embedding vectors with d dimension ($\boldsymbol{z}_e^{(i,j)} \in \mathbb{R}^d$, $i \in \{1, 2, \ldots, W\}$, $j \in \{1, 2, \ldots, H\}$). Next, $H * W$ embedding vector of latent representation $\boldsymbol{z}_e(\boldsymbol{x})$ is further mapped to a discrete latent matrix $\boldsymbol{z} \in \mathbb{R}^{H*W}$. Here, each $z^{(i,j)} \in \boldsymbol{z}$ is the index of the nearest concept $\boldsymbol{c}_{\mathrm{nrs}}$ in the codebook for each $z_e^{(i,j)}$ via nearest neighbor searching $\arg\min_m \left\| z_e^{(i,j)}(\boldsymbol{x}) - \boldsymbol{c}_m \right\|$, i.e., the Vector Quantization. The decoder reconstructs the concepts in the pixel space by using the quantized embedding items $z_q(\boldsymbol{z})$ corresponding to the discrete latent concept index matrix via another non-linear function. Trainable components are the

encoder, decoder, and codebook. The loss function is:

$$\mathcal{L}_{\text{VQ-VAE}} = \text{Dist}(\boldsymbol{x} - \text{De}(\boldsymbol{z}_q(\boldsymbol{z}))) + \|\text{sg}[\text{En}(\boldsymbol{x})] - \boldsymbol{c}_{\text{nrs}}\|_2^2 + \beta\|\text{sg}[\boldsymbol{c}_{\text{nrs}}] - \text{En}(\boldsymbol{x})\|_2^2, \quad (1)$$

where $\text{sg}[\cdot]$ is a stop-gradient operation that blocks gradients from flowing into its argument, β is a hyperparameter, and $\text{Dist}(\boldsymbol{x} - \text{De}(\boldsymbol{z_q}(\boldsymbol{z})))$ reveals the reconstruction error. The procedures of training the encoder and decoder, updating the codebook and configuration of the exponential moving average follows two works [55] and [47].

2.2.2 Concept Alignment

To boost the utilization of concept codebook, we apply two enhancement strategies: clustering-based alignment and distance-based similarity to find the nearest vector.

Clustering-based Alignment. Given a pre-trained encoder and a learned codebook, the high-dimensional input can be transformed into an index matrix of concepts within the codebook. However, evaluations (demonstrated in Appendix B) showed that not all K concepts of the codebook were utilized, limiting its effectiveness for subsequent diagnostic procedures. A clustering-based alignment for the codebook mitigates this under-utilization. After the first round of training, all image embeddings $\boldsymbol{z}_e$ are extracted from the training set using the current version of the encoder $\text{En}(\cdot)$ and codebook $C = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_K\}$. Next, cluster centers are computed with k-means ($k = K$) for the set of $\boldsymbol{z}_e$. These new centers replace the existing concepts in the codebook. Finally, we fine-tune the VQ-VAE on the training data as the second round training. This ensures a reversible mapping between the updated codebook and images, enhancing the effectiveness of the diagnostic processes (improvements displayed in Appendix B).

Distance-based Similarity. To assist the selection of the nearest quantized concept for each embedding $\boldsymbol{z}_e(x)$, we incorporate the similarity between $\boldsymbol{z}_e(x)$ and its nearest concept as an additional feature for the index. Intuitively, the Euclidean distance between the pair serves as a metric for evaluating similarity. However, the Wasserstein distance, also known as the Earth Mover's distance, has proven to be effective in the context of generative models and continuous representation learning [6,22,53]. The advantage of the Wasserstein distance is that it can still reflect the closeness of two distributions even if the support sets of the two distributions do not overlap or have very little overlap. Therefore, we employ the Wasserstein distance as a similarity metric to align the embedding $z_e(x)$ to concepts in the codebook. Smaller Wasserstein distances indicate higher similarity between two vectors.

2.3 Diagnosis of DNNs' Faults

For diagnosing a DNN fault within a given sample, we rely on the **Index Matrix** $\boldsymbol{Z}$ and the **Distance Matrix** $\boldsymbol{Z_d}$. The index matrix captures the indices of the closest concepts from the codebook for each embedding vector in the input's latent representation, denoted as $\boldsymbol{z}_e(x)$. Concurrently, the distance matrix chronicles the alignment distances between these embedding vectors and their matched concepts. Utilizing both $\boldsymbol{Z}$ and $\boldsymbol{Z}_d$, we derive a composite feature vector for each sample by amalgamating the two matrices and integrating essential statistics (such as mean, variance, maximum, minimum, etc.) from the distance matrix. Subsequently, the final *basic feature* vector is generated through a Principal Component Analysis (PCA) [1] applied to this composite feature vector.

$$\boldsymbol{Z} = \begin{bmatrix} i_{11} & \dots & i_{1\text{H}} \\ \vdots & \ddots & \vdots \\ i_{\text{W}1} & \dots & i_{\text{WH}} \end{bmatrix}, \boldsymbol{Z_d} = \begin{bmatrix} d_{11} & \dots & d_{1\text{H}} \\ \vdots & \ddots & \vdots \\ d_{\text{W}1} & \dots & d_{\text{WH}} \end{bmatrix} \quad (2)$$

Pipeline. Our diagnostic investigation centers on the hypothesis: *"The alignment index matrix and alignment distance matrix can effectively differentiate attack samples from the original ones?"* If true, the subsequent question becomes: *"What malicious patterns emerge from this differentiation?"* Diagnostic exploration is structured into three processes:

***(i)* Kolmogorov-Smirnov (KS) Test.** We first evaluate the *Null Hypothesis $\mathcal{H}_0$: the samples of the original and attack samples are drawn from the same distribution* (Section 2.3.1).

***(ii)* Distance Matrix Diagnosis.** Next, we scrutinize discerning patterns or discrepancies from the sample-wise and position-wise distance distributions of the raw alignment matrix $\boldsymbol{Z}_d$ between original and attack. If the discrepancies remain elusive using raw $\boldsymbol{Z}_d$, we transition to the spectral domain of $\boldsymbol{Z}_d$, aiming to amplify underlying patterns (Section 2.3.2). The rationale behind spectral analysis is its efficacy in revealing structures and patterns which may not be immediately evident in the spatial domain.

***(iii)* Concept Index Diagnosis.** We subsequently examine the aligned concept index matrices of both the original and attack samples. Through a thorough assessment of the distribution and prominence of latent concepts, our objective is to pinpoint specific concept patterns that consistently correlate with faulty samples (Section 2.3.3).

***(iv)* Transition Patterns Diagnosis via Reconstruction.** We compare the investigating samples to their reconstructed counterparts in the pixel space. This diagnostic approach permits visual comparison to transfer nonstructural perturbation to malicious structural patterns, as depicted in Figure 1(c).

2.3.1 Kolmogorov-Smirnov (KS) Test

The KS test [40] is a non-parametric hypothesis test to determine if two sample distributions are different. The KS test operates under the *null hypothesis $\mathcal{H}_0$: the samples are drawn from the same distribution*. Given two empirical cumulative distribution functions (ECDFs), $F_n(x)$ and $G_m(x)$, the KS statistic D is defined as $D = \max_x |F_n(x) - G_m(x)|$. Intuitively, this statistic quantifies the greatest vertical distance

between the two ECDFs. Two values are calculated for the hypothesis testing, statistic D and p-value. The test statistic D is the maximum difference between the two cumulative distributions. The p-value is the probability of observing the test statistic as extreme as, or more extreme than, the statistic computed from the sample under the null hypothesis. Small p-values (typically $p < 0.05$) reject the null hypothesis.

2.3.2 Distance Matrix Diagnosis

We utilize three metrics for evaluating the distance matrix: sample-, position-wise distances, and spectral properties.

Sample-aware Distance Metrics. Given the distance matrix for a sample, we calculate the average distance across all position of the matrix: $d_s = \frac{\sum_i^W \sum_j^H d_{\text{ij}}}{H * W}$. Then, we calculate statistics like the mean and variance for all d_s of the dataset.

Position-aware Distance Metrics. We aggregate each position value of the distance matrix $\boldsymbol{Z}_d$ over all samples in the dataset: $d_p^{(i,j)} = \frac{\sum_x d_{\text{ij}}}{|\boldsymbol{X}|}$, where $|\boldsymbol{X}|$ is the size of the dataset. We also calculate the statistics of distance distribution on each position, like mean and variance. The variance provides insights into which specific areas of the latent representation deviate more than others.

Spectral Characteristic Metrics. Further, by applying the Fourier Transform [10] to the distance matrix, we can analyze the spectral properties (frequency components and energy intensity) of the distance values. With a quantification of energy across frequency components, we can discern potential anomalies in data between original and attack. We apply 2D Fourier Transform to transform the 2D distance matrix $\boldsymbol{Z}$ to the frequency domain: $F(u,v) = \sum_{i=1}^{W} \sum_{j=1}^{H} f(i,j) e^{-2\pi\left(\frac{u_i}{W} + \frac{v_j}{H}\right)}$, where $f(i,j)$ is an element of the distance matrix $\boldsymbol{Z}_d$. u and v represent spatial frequencies in the horizontal (along the x-axis) and vertical (along the y-axis) directions, respectively. To better visualize frequency components, the zero-frequency component is shifted to the center. When capturing spectral information, low frequencies typically represent principal information, while higher frequencies typically pertain to noise or finer details. Filters are used to isolate specific frequencies. The intensity of a frequency (and thus the information component) is proportional to its energy $E = \sum |F(u,v)|^2$.

2.3.3 Index Matrix Diagnosis

To emphasize the indices undergoing the most frequent changes or increased usage, we compare pairs of concept index metrics between original and attack data.

Concept Distribution. The usage of indices may reveal specific patterns unique to the original or attack samples.

Frequently Changed Position. Given a paired index matrix for an original sample x and its attack or faulty counterpart x', we compute a Boolean difference matrix, $\boldsymbol{BD}$, to discern discrepancies across the positions of both index matrices. Specifically, $BD^{(i,j)}$ is set to 0 if $\boldsymbol{Z}_x^{(i,j)} = \boldsymbol{Z}_{x'}^{(i,j)}$, otherwise 1. Accumulating these differences across all samples pinpoints positions undergoing concept changes most frequently. We spotlight the top-20 positions with the most recurrent shifts from the original to the adversarial samples, referred to as 'hotspots'. This offers insights into potential vulnerabilities or areas particularly prone to perturbation.

Transition Patterns. We examine the transition patterns between original and adversarial samples for individual positions. This analysis captures the most prevalent transition pairs $(\boldsymbol{Z}_x^{(i,j)}, \boldsymbol{Z}_{x'}^{(i,j)})$ and computes the entropy for each transition originating from a particular original index. The entropy for a concept index s is given by: $H(s) = -\sum_t p(s,t) \log_2 p(s,t)$, where $p(s,t)$ represents the probability of the transition pair $(\boldsymbol{Z}_x^{(s,t)}, \boldsymbol{Z}_{x'}^{(s,t)})$ among all possible transitions from s. Indices characterized by high entropy indicate diverse transitions, suggesting unpredictability. In contrast, those with low entropy signal consistent transitions to specific adversarial indices, offering potential insights into predictable adversarial strategies.

2.4 Mitigation by Detection

Beyond diagnosis, we construct detection approaches to distinguish between the class of malicious/abnormal samples and the original samples. Our primary objective in mitigation is to "Assess the richness of latent conceptual representations for the downstream tasks". To this end, we employ a simple end-to-end machine learning model for detection, leveraging our *basic features* (see Section 2.3) as a baseline. Only features from the original dataset is used to train an unsupervised detector, which is tested against our multiple fault settings with different attacks. We shall apply the unsupervised method of One-Class SVM [49] later in Section 6. We acknowledge other unsupervised models exist, including Local Outlier Factor (LOF) [9] and Elliptic Envelope [48], however One-Class SVM was sufficient to achieve state of the art results, and used as the lower bound of the effectiveness.

One-Class SVM finds a hyperplane that optimizes the separation between data points considered normal, with points outside this hyperplane as anomalies. One-Class SVM does not make strong assumptions about the requirement of the clean (or "normal") data. Given a set of training data, the objective of a one-class SVM is to find a function f such that: $f(x) = \mathbf{w} \cdot \phi(\mathbf{x}) - \rho$, where $\mathbf{w}$ is the weight vector, ϕ is a function mapping data to a higher-dimensional space (if a kernel trick is used), and ρ is a threshold. Samples with $f(\mathbf{x}) \leq 0$ are considered anomalies.

Metrics used to measure model performance include: Accuracy for a model's ability to correctly identify both faulty and original samples, and False Positive Rate (FPR) captures the ratio of mislabeled faulty samples as original samples.

3 Evaluation Settings

Our analysis compares diagnoses from an *Original Set* and an *Attack Set*. The Original Set is randomly sampled from the original validation set and serves as a point of reference. The Attack Set contains samples that are expected to exhibit notable change in the output of target models after perturbation, transformation or data drift. The following evaluation seeks to answer the following questions: *"Do the index and distance matrices contain enough distinguishing information to distinguish between original and attack (adversarial, semantic, or drifted) samples?"* and *"Do attack samples exhibit unique behavioral patterns within the index and distance matrices?"*

Datasets. Noting the increased difficulty of concept extraction from images compared to tabular data, in our experiments, we employ three benchmark datasets found in image classification tasks, including MNIST [31], CIFAR-10 [30] and CelebA [34]. Dataset specifics are expanded in Appendix C.

Models. We utilize Convolutional Neural Networks (CNNs), e.g., ordinary CNN with different layers, and standard models of ResNet-18 [23] and DenseNet-121 [26] as target models of evaluation. Detailed configurations are in Appendix C.

Attacks. We study adversarial attacks of FGSM, PGD, JSMA, DeepFool, CW attacks, and Pixel Attack. For semantic attacks, we investigate rotation and color shift attacks. The detailed description of such attacks can be found in Appendix A.

Detection Benchmarks. We use existing detection benchmarks for comparison, including unsupervised {Z-Score [52], NIC [37], MagNet [41](reconstruction error-based), and supervised LID [38]. Default settings are same with [4].

Difference. We reiterate that DNN-GP's detection approach delineates itself from other benchmarks through several key attributes. DNN-GP operates unsupervised and requires no white-box information of a model. Furthermore, it is lightweight and relies exclusively on the original data. Notably, its design is both fault-agnostic and model-independent.

For evaluating adversarial attacks, given the limited efficacy of current methodologies, we offer an extensive evaluation of the CIFAR-10 results in the main text, framing it as a worst-case scenario. While we provide brief summaries for the MNIST and CelebA datasets within the main body, their detailed results are presented in Appendix D. For semantic attack evaluation, we consider rotation for MNIST, CIFAR-10 and CelebA, contrast for MNIST, and Color shift for CIFAR-10 and CelebA. Our subsequent evaluation is separated by the type of fault, with Section 4 diagnosing Adversarial Examples, Section 5 diagnosing Semantic Attacks, and Section 6 evaluating downstream tasks of detection for attacks and data drift. Appendix D.2 presents auxiliary diagnosis of Data drift with similar conclusions to semantic attacks. Default (K,d,H,W) values for the VQ-VAE are given as: $(512,40,7,7)$ for MNIST, $(512,64,8,8)$ for CIFAR-10, and $(512,64,64,64)$ for CelebA. Default values for the One-Class SVM are: nu $= 0.04$, kernel = "rbf", gamma $= 0.01$.

4 Diagnosis of Adversarial Examples

In this section, we perform a detailed diagnosis of perturbations that induce adversarial attacks. Our objective is to decipher the mechanics underpinning these perturbations and comprehend their conceptual implications. With foresight from our reconstruction error-driven detection results later in Section 6, we categorize the perturbations by stealthiness: ***(i)* Obvious Perturbations:** These are typically associated with attacks like FGSM or PGD that leverage larger perturbation factors (e.g., $\varepsilon \geq 32/255$). ***(ii)* Minimal Perturbations:** Bounded perturbations like CW, or those utilizing smaller scales (e.g., $\varepsilon < 32/255$). ***(iii)* Pixel Attack Perturbations:** Attacks whereby perturbations change fewer than three pixels.

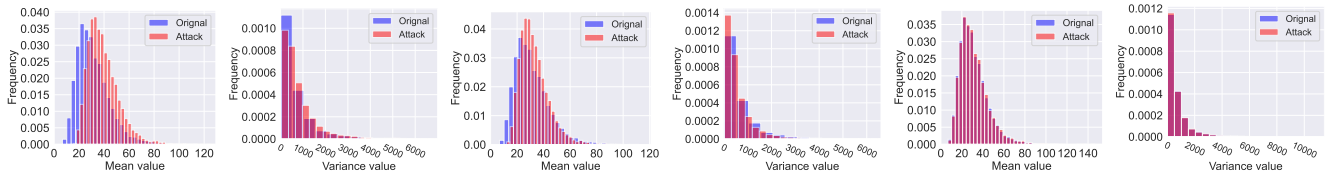
4.1 Kolmogorov-Smirnov (KS) Test

To demonstrate the insights from the KS test, we first use FGSM with noise magnitude $32/255$ (obvious perturbation), CW-inf (minimal perturbation), Pixel (a challenging case), these tree form the basis of our evaluation moving forwards. We additionally include PDG, Deepfool and JSMA for additional comparisons. Table 1 presents the results of the KS test comparing basic feature distributions between the original and attack CIFAR-10 datasets. A key observation is that every adversarial attack type produces a p-value substantially below 0.05 across all noise levels. This outcome rejects the null hypothesis $\mathcal{H}_0$, suggesting that the original and perturbed data do not share the same distribution. Furthermore, as the magnitude of adversarial noise intensifies, there is a notable escalation in the D-statistic, leading to an even more diminished p-value. For example, as the noise magnitude in FGSM varies from $8/255$ to $32/255$, the D-statistic surges from 0.002 to 0.036. In each instance, the p-values remain below the 0.05 threshold. This suggests that higher adversarial noise magnitudes result in a distribution increasingly distinct from the original. Importantly, these significant distributional disparities arise regardless of the subtlety of the adversarial perturbations. The results on MNIST and CelebA also confirm these findings, with p-values substantially below 0.05 across all attack types and noise levels. We shall now delve into the patterns contributing to discrimination from the alignment distance matrix and index matrix.

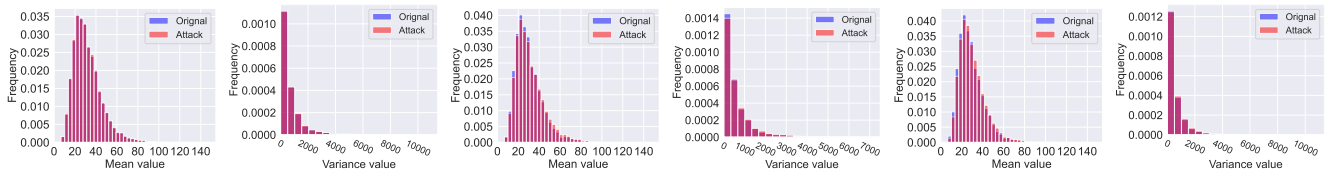
4.2 Distance Matrix Diagnosis

Our diagnostic method for the distance matrix adopts two perspectives: one focuses on individual samples, while the other emphasizes particular positions within the matrix.

Sample-aware Metrics. The mean and variance of sample-aware distance between original and attack datasets is depicted in Figure 3. Seen from Figure 3a, we observe attacks with obvious perturbations (e.g., FGSM-32/255) demonstrate a distinct histogram pattern, with clearly separable peaks. We



(a) Mean (FGSM-32) (b) Var. (FGSM-32) (c) Mean (PGD-32) (d) Var. (PGD-32) (e) Mean (JSMA, inf) (f) Var. (JSMA, inf)



(g) Mean (DF, inf) (h) Var. (DF, inf) (i) Mean (CW, inf) (j) Var. (CW, inf) (k) Mean (Pixel, 3) (l) Var. (Pixel, 3)

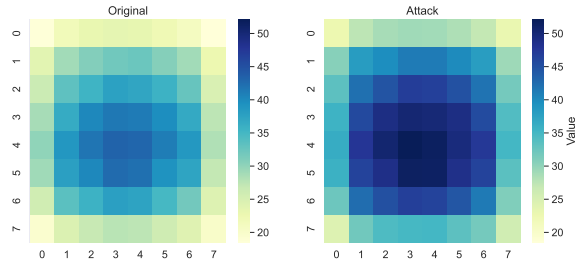
Figure 3: Sample-wise distance distributions of original and adversarial samples for different adversarial attack methods.

Table 1: KS test results between different attacks and the original CIFAR-10 distribution.

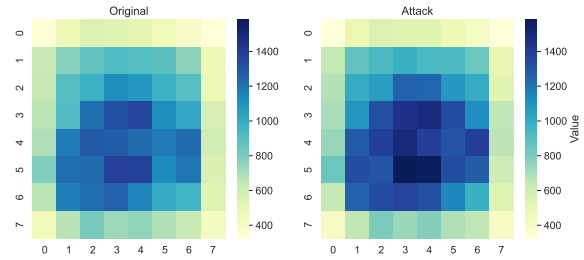
Attack Setting		D-value	p-value
Attack Method	Noise Parameter		
FGSM	8/255	0.002	0.048
	16/255	0.036	4.45×10^{-5}
	32/255	0.036	1.84×10^{-23}
PGD-Linf	8/255	0.008	2.49×10^{-17}
	16/255	0.029	1.30×10^{-219}
	32/255	0.089	≈ 0
CW-Linf	-	0.030	0.008
Pixel	3	0.003	0.030
DeepFool	-	0.002	0.022
JSMA	-	0.008	7.00×10^{-18}

also observe the variance of the informed datasets exceeds that of the original, signaling a broader distribution around the mean. When it comes to minimal perturbations (Figure 3i, CW) or pixel perturbations (Figure 3k, pixel attack with 3 pixels change), the distance distributions overlap. When the perturbation is small, attack samples are difficult to separate from the original samples with only the mean or variance of sample-wise distance.

Position-aware Metrics. Delving into the position-wise mean and variance metrics in the latent concept distance matrix visualized in Figure 4, the heatmaps provide a spatial sense of changes among the positions in a sample. We generally see larger perturbations yield relatively increased means, echoing our earlier observation in the sample-aware distributions. For the minimal perturbation scenario, the proportion of positions with pronounced distance values is less than the obvious perturbation. Nevertheless, examining the position-wise variation serves as an effective method to identify which positions are more susceptible or resilient.



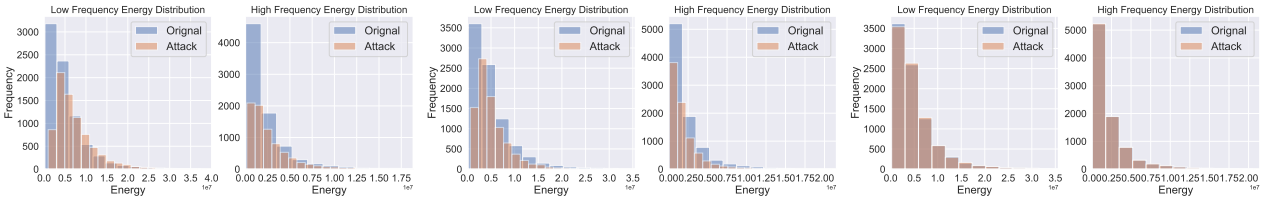
(a) Mean (FGSM-32)



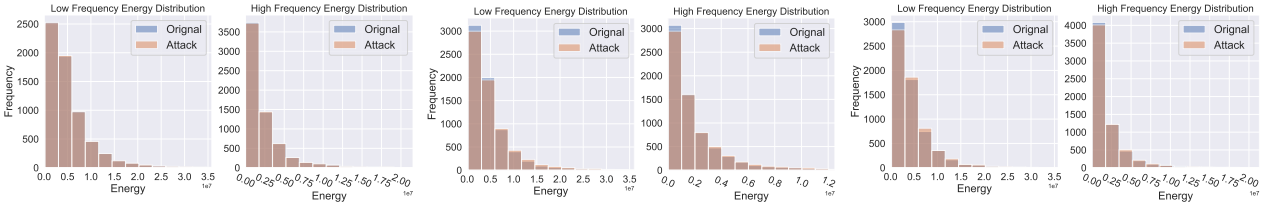
(b) Variance (FGSM-32)

Figure 4: Position-wise distance distributions of original and adversarial samples with the FGSM attack.

Spectral Characteristic Metrics. Comparing the energy distribution of the low and high-frequency components between the original and attack data in Figure 5 observes differences in both low and high-frequency energy distributions between the original. The obvious perturbations have significant impacts on both overall structure (high-frequency) and detailed information (low-frequency), while minimal and pixel attacks mainly affect the detailed information and have no impact on the overall structure of the object. The high-frequency components can also capture some features of the affected areas. We also find that the influence of the perturbation is mainly reflected in the high-frequency components. Figure 6 confirms the spatial patterns in the respective high frequency

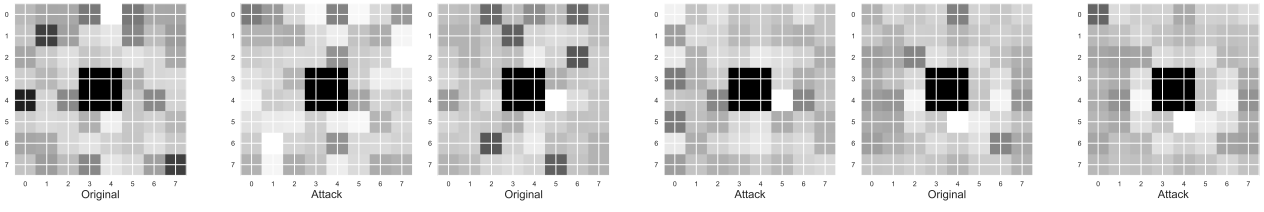


(a) Low/High energy dist. (FGSM-32) (b) Low/High energy dist. (PGD-32) (c) Low/High energy dist. (JSMA, inf)

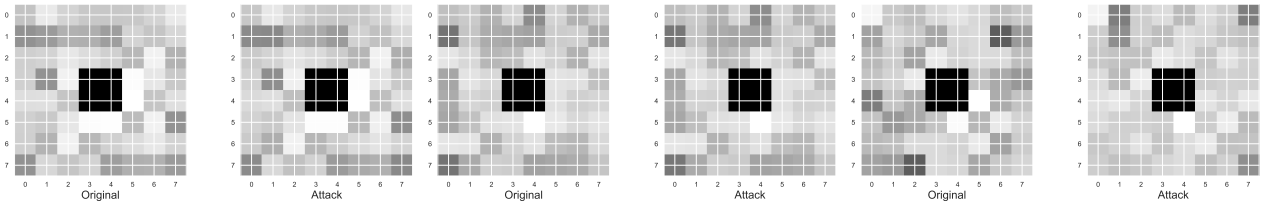


(d) Low/High energy dist. (DF, inf) (e) Low/High energy dist. (CW, inf) (f) Low/High energy dist. (Pixel, inf)

Figure 5: High/low spectral energy distributions of original and adversarial samples for different adversarial attack methods.



(a) Dominant high frequency (FGSM-32) (b) Dominant high frequency (PGD-32) (c) Dominant high frequency (JSMA, inf)



(d) Dominant high frequency (DF, inf) (e) Dominant high frequency (CW, inf) (f) Dominant high frequency (Pixel, inf)

Figure 6: Dominant high frequency plots of original and adversarial samples for different adversarial attack methods.

components. The high frequency components magnify the discrepancy compared to the raw distance matrix domain.

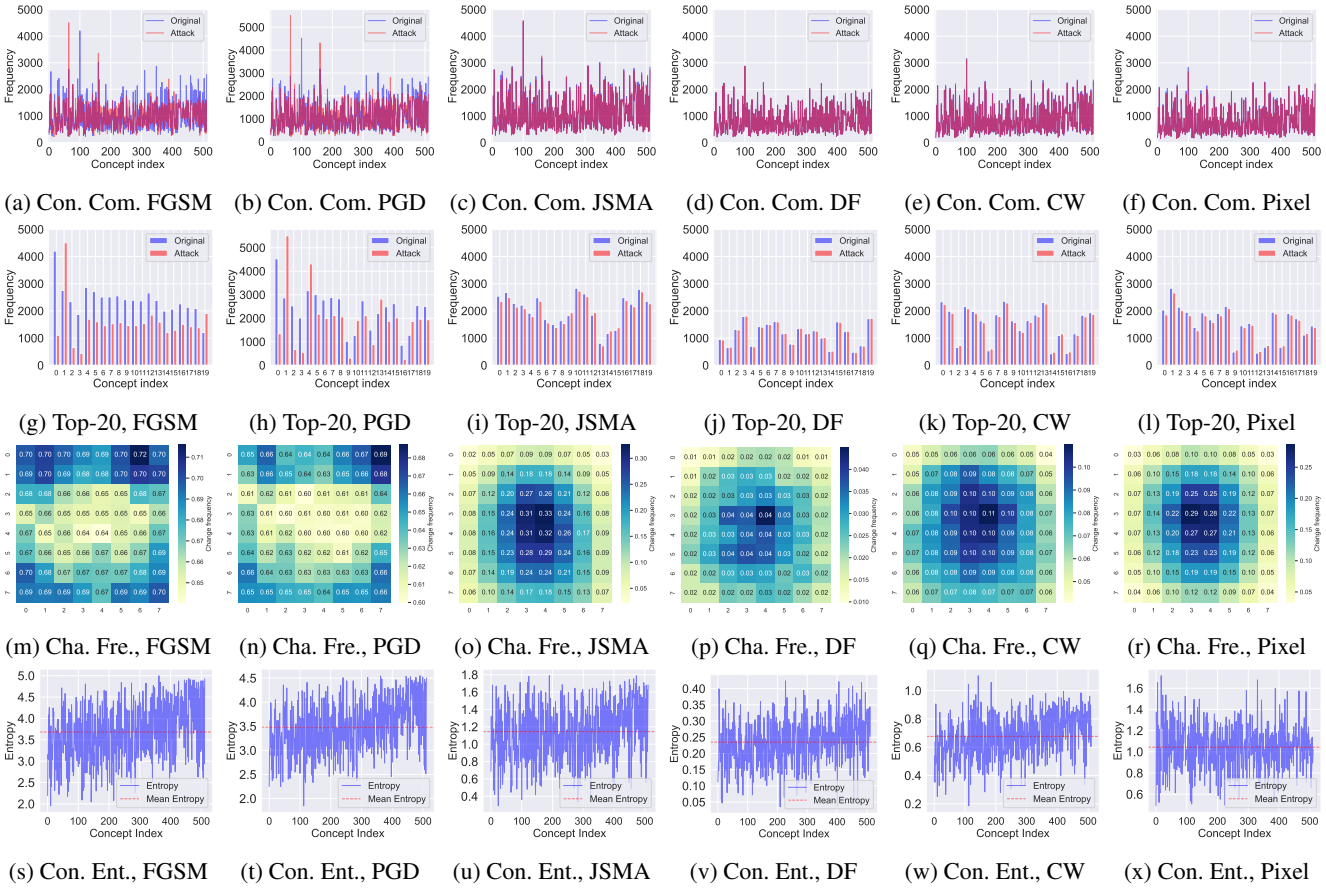
These results provide a compelling visual representation of the differences in the distributions of distance. It is evident that the adversarial perturbation exhibits a distinct distribution pattern when compared to their original counterparts. Additionally, prominent perturbations exert a widespread influence across the data. In contrast, minimal and pixel-specific attacks are meticulously crafted to minimize their impact on the low-frequency components, focusing their perturbations predominantly on the high-frequency components.

We additionally present the efficacy of DNN-GP for DenseNet-121 and ResNet-18 architectures on CelebA-10, with comparable conclusions drawn from results shown in Appendix Figure 12 and 13. In conclusion, for obvious perturbations, differentiation can be accomplished by solely relying on alignment distance. Conversely, reliance on alignment distance alone presents challenges when differentiating minimized perturbations. This necessitates in-depth exploration of the features within the alignment index.

4.3 Index Matrix Diagnosis

The overall frequency and alterations of latent concepts occurring between the original and attack samples in addition to the spatial distribution of changes and the entropy of concepts are visualized in Figure 7.

Concept Distribution Analysis. We fist show the frequency distribution of specific concept indices. The first row of Figure 7 delineates the usage and prevalence of certain indices for both the original and adversarial sets. Identifying the most recurrent indices in both the original and attack sets offer



(a) Con. Com. FGSM (b) Con. Com. PGD (c) Con. Com. JSMA (d) Con. Com. DF (e) Con. Com. CW (f) Con. Com. Pixel

(g) Top-20, FGSM (h) Top-20, PGD (i) Top-20, JSMA (j) Top-20, DF (k) Top-20, CW (l) Top-20, Pixel

(m) Cha. Fre., FGSM (n) Cha. Fre., PGD (o) Cha. Fre., JSMA (p) Cha. Fre., DF (q) Cha. Fre., CW (r) Cha. Fre., Pixel

(s) Con. Ent., FGSM (t) Con. Ent., PGD (u) Con. Ent., JSMA (v) Con. Ent., DF (w) Con. Ent., CW (x) Con. Ent., Pixel

Figure 7: Latent concept index matrix metrics, *Con. Com.* represents 'concept comparison', *Top-20* refers to the top-20 indices exhibiting the largest discrepancies between the original and adversarial samples, *Cha. Fre.* refers to 'change frequency', and *Con. Ent.* refers to 'concept entropy'. The attacks are configured as follows, FGSM as (FGSM, 32/255), PGD as (PGD, 32), JSMA as (JSMA, inf), DF as (DeepFool, inf), CW as (CW, inf), and Pixel as (Pixel, 3).

insight into specific attack patterns or abnormalities. For example, Index 498, which corresponds to a "black background" upon reconstruction in the original dataset, emerges as a favored target for adversarial manipulation. In the second row of Figure 7, we emphasize the top-20 indices that exhibit the greatest discrepancy between the original and adversarial samples, highlighting instances where certain indices are notably more prevalent in one set over the other. For obvious perturbations, a strong contrast is evident in the frequently used indices between the original and adversarial samples. Both minimal and pixel perturbations exhibit a frequency distribution more akin to the original than the obvious perturbations do, with only slight differences in index usage between the original and the minimal/pixel perturbations. This underscores that obvious perturbations might introduce more distinct conceptual patterns than minimal perturbations.

Frequent Change Positions. We now investigate how faults induce changes to the latent concepts with regards to the relative positions of the latent concepts. Darker spots on the third row heatmaps of Figure 7 indicate positions with larger difference exist, implying that these positions are more frequently changed under an adversarial attack. The obvious perturbations indicate high amounts of conceptual change in across all positions (over 82% in all datasets). For the minimal and pixel perturbations, the greatest values of changes are lower, and are found towards the center of the concept heatmap. All of these perturbations a high frequently of changed positions, indicating that these positions are the vulnerable positions to be targeted. As certain positions within the grid exhibited more pronounced changes than other, recognizing these regions can be crucial for understanding and locating vulnerabilities or areas most affected by the attack.

Transition Patterns. We further calculate the entropy for each latent concept from the frequency of change in the sample sets. Low entropy reflects consistency in concepts between original and attack samples, with high entropy indicating greater transition between concepts. Consequently, concepts with high entropy values are more susceptible to being changed during the attack. The fourth row of Figure 7 shows pronounced perturbations yield a higher average entropy than minimal

and pixel perturbations. This can be attributed to the tighter constraints imposed on the design of the latter perturbations. Furthermore, it is feasible to identify a collection of recurring transition patterns at positions that garner significant interest.

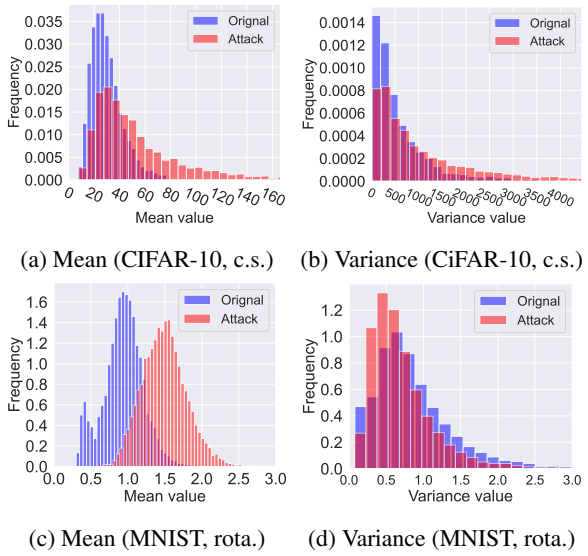
4.4 Pixel-space Diagnosis

We also analyze adversarial perturbations based on reconstruction. Understanding the nature of the visual perturbations can provide insights into the conceptual vulnerabilities of the model and potential strategy of the attacker. In scenarios where the perturbation is evident to the human eye, the reconstructed versions of the attacked images can shed light on the meaning of latent concept the attacker attempts to exploit. For example, the absence of a small stroke or strike of the digit in Figure 1, with more examples in Appendix D.3. Such structural patterns may not be revealed in the attacked image only becoming evident after reconstruction. This suggests that the adversarial attack introduces features that sufficiently changes key latent concepts that alter the target class. Further, the adversarial output and reconstruction indicate this latent concept contains a vulnerability linking it to image regions beyond the visible stroke. On the other hand, vector quantization plays a pivotal defensive role by representing the original data with fewer bits, with similar data points mapped to the same concepts in the latent space. In the context of adversarial attacks, this means that the subtle, malicious changes introduced by an attacker can be"averaged out" or reduced during the quantization process. Thus, when the attacked image undergoes vector quantization and subsequently reconstructed, the impact of the adversarial perturbation is diminished, akin to a "purification" process. Despite these reconstructions mitigating perturbations, they retain and emphasize the essential structural and content influences of the original image. This dual characteristic ensures that while the adversarial noise is filtered out, the salient features of the image, which may have been the target of the adversary, remain discernible. Such a process offers valuable insights into both the robustness of the reconstruction mechanism and the specific targets or vulnerabilities that adversaries seek to exploit in a given model.

5 Diagnosis of Semantic Attacks

Recall that our objective remains to understand influence a semantic attack has on latent concept. Specifically, we investigate semantic attacks including rotation, contrast change on MNIST and the color shift attack with CIFAR-10.

KS Test. The KS test for rotation, contrast and color shift yields a D statistic less than 0.05 with a p-value less than 0.05. As such we can confidently reject the null hypothesis that the samples of original and semantic attack are drawn from the same distribution.



(a) Mean (CIFAR-10, c.s.) (b) Variance (CiFAR-10, c.s.)

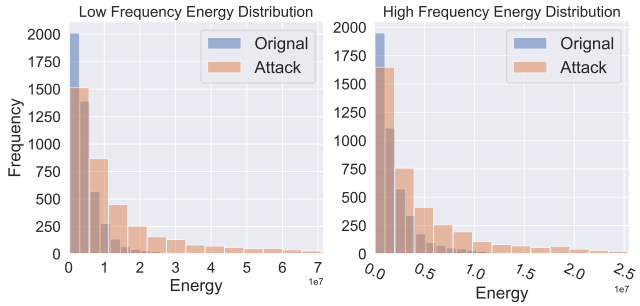
(c) Mean (MNIST, rota.) (d) Variance (MNIST, rota.)

Figure 8: Sample-aware distance distributions of original samples and semantic attack samples. Here, *c.s.* indicates color shift attacks, and *rota.* indicates rotation attacks.

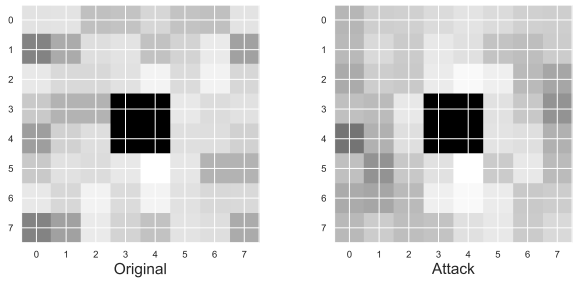
5.1 Distance Matrix Diagnosis

Sample-aware Distance Metrics. Recall the position-aware distance metrics are derived from the difference between the sample embedding vector and our codebook concepts. The mean and variance of these deviations are computed across the samples of the original (blue) and attack (red) data in Figure 8. There is an observable difference between the mean peaks, albeit, with overlap. Specifically, average sample-wise distance values show differences that separate the original and attack distribution. Their variance is also spread across a broader range for rotation and contrast change, emphasizing its increased variability. Smaller scales of the transformation in smaller distributions of distances are yielded by spatial rotation; however notably, the peaks of the distribution are clearly separable. In all attacks, the attack distributions skew to the larger side, presenting concepts that are unlike naturally aligned codebook concepts.

Spectral Characteristic Metrics. From the low and high frequencies components shown in Figure 9a, the low-frequency component can capture the overall structure of the object, with color shift having no significant impacts on the overall structure of the object. The high-frequency components of Figure 9b also capture contextual features of the color shift with similar observations for rotation and contrast in Appendix D. By comparing the energy distribution of the low and high-frequency components between the original and semantic attack data, we can see that there are significant differences in high-frequency energy distributions between them while the low-frequency parts remain the same for all three semantic attacks.

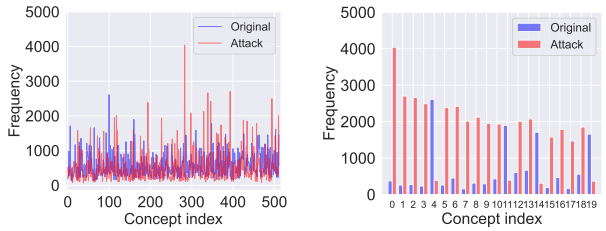


(a) Low/High energy dist.



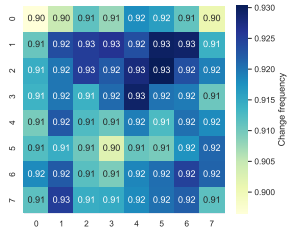
(b) Dominant high frequency

Figure 9: Spectral characteristics of distance matrices between original and attack samples from the color shift attack.

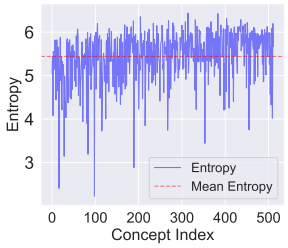


(a) Concept comparison

(b) Top-20 comparison



(c) Change frequency



(d) Concept entropy

Figure 10: Latent concept index matrix metrics for the color shift semantic attack.

5.2 Index Matrix Diagnosis

Concept Distribution Analysis. With the histogram of frequency of latent concepts between our original and semantic attack samples displayed in Figure 10a, we can identify if particular concepts are more commonly to attack than others. We observed more unique index values that only appeared in either original or semantic attack samples, albeit in even lower quantities (in contrast to adversarial attacks). It is possible to observe the clear difference in the top-20 latent concepts with the largest differences after the semantic attack in Figure 10b.

Frequent Change Positions. Figure 10c's heatmap illustrates the position-wise spatial patterns between the original and attack concept index matrices. Brighter spots imply that these positions are more frequently altered during attacks.

We observe semantic attacks cause general changes of the aligned index for most positions in contrast to adversarial attacks. The widespread changes are a result of the transformation incorporating structural pattern changes that affect many pixels. Even with the color shift attack's global shift, there are still specific positions of the index matrix that retain low change frequencies. As such it is possible to design robust defenses or detection using these positions.

Transition Patterns. We finally calculate the entropy of transitions for each original index based on the original to attack sample transition matrix in Figure 10d. Indices with high entropy suggest that there is much variability in how they transition to attack indices, while those with low entropy tend to transition in a more consistent and predictable manner to specific attack indices. Additionally, some transition pairs have higher intensities, indicating certain transitions that occurred more frequently. This suggests that there are specific latent values are more susceptible to being changed during an attack, useful for both detection and mitigation efforts.

5.3 Pixel-space Diagnosis

Our final diagnosis delves into the transformations achieved through a decoder-based reconstruction for semantic attacks. We briefly describe our findings, with visual demonstrations of these transformations shown in Appendix D.3. Across all transformations, the reconstructed versions of the compromised images illuminate the intrinsic patterns targeted by the attacker. This reconstruction-based analysis is particularly beneficial as it demystifies the subtle adversarial alterations of each specific transformation, revealing the regions and features of the image that attackers deem critical for successful adversarial manipulation. By comprehensively examining these reconstructed images, researchers and practitioners alike can gain insights into potential areas of improvement for model robustness against semantic attacks.

6 Mitigation Countermeasures: Use Cases

We first benchmark DNN-GP's performance against standard detectors, followed by DNN-GP evaluated under adaptive attacks and data drift.

Table 2: Adversarial attack detection performance on CIFAR-10. DR and FPR is respectively the Detection Rate and False Positive Rate of a given detection method.

Attack Setting		Detection Baseline								**DNN-GP (Ours)**	
Attack type	Noise parameter	MagNet [41]		Z-score [52]		NIC [37]		LID [38]		One-Class SVM [49]	
		DR	FPR	DR	FPR	DR	FPR	DR	FPR	DR	FPR
FGSM [20]	8 / 255	0.07	0.045	0.25	0.219	0.436	0.101	0.54	0.315	1.00	0.04
	16 / 255	0.453	0.039	0.266	0.219	0.96	0.101	0.712	0.009	1.00	0.04
	32 / 255	1	0.039	0.469	0.219	0.995	0.101	0.915	0.001	1.00	0.04
PGD-Linf	8 / 255	0.065	0.044	0.188	0.219	0.834	0.101	0.649	0.004	1.00	0.04
	16 / 255	0.237	0.046	0.219	0.219	0.961	0.101	0.795	0.027	1.00	0.04
	32 / 255	1	0.046	0.25	0.219	1	0.101	0.96	0.011	1.00	0.04
CW-Linf [14]	-	0.233	0.039	0.313	0.219	0.951	0.101	0	0	1.00	0.04
Pixel [29]	3	0.046	0.04	0.25	0.234	–	–	0.741	0.252	1.00	0.04
Deepfool	-	0.05	0.05	0.25	0.25	0.919	0.949	0.834	0.101	0.998	0.04
JSMA	-	0.058	0.046	0.234	0.219	–	–	0.846	0.065	0.999	0.04

6.1 Detection Performance

We employ the detector detailed in Section 2.4 as a mitigation countermeasure, the detection results for CIFAR-10 are consolidated in Table 2. Notably, our proposed detector surpasses all pre-existing methodologies. Test samples consist of 500 successful attack samples based on 500 randomly selected testing original samples (100 from 100 for CelebA). Across all adversarial perturbations, from obvious (FGSM), minimal (CW), Deepfool, JSMA, pixel attacks, and black-box attacks (e.g., Square Attack) on CIFAR-10, our detector, leverages basic feature vectors in tandem with a One-Class SVM to achieve a detection rate of 100% and a false positive rate of 4%. This places DNN-GP significantly ahead of the current state-of-the-art methods, including NIC.

The detected 4% false positive rate within the original dataset indicates that a small subset of genuine data points was mistakenly classified as adversarial, as the cost of one-class classification. This rate can be further improved through sample reconstruction techniques, presented in Table 4 of Appendix D.3. Through reconstruction, the original samples observe a minimal accuracy reduction of 2% when transitioning to their reconstructed versions within the same classifier framework. Remarkably, for adversarial samples, over 87% are rectified to their correct labels, effectively neutralizing the malicious perturbations.

6.2 Analysis of Adaptive Attacks

In light of emerging threats from adaptive adversaries, devising a robust defense strategy necessitates a deep understanding of potential attack avenues. We examine two potent cases:

A1: Leveraging DNN-GP diagnosis for Adversarial Guidance. Attackers may employ the diagnostic insights from DNN-GP to identify and target specific regions that are either robust or susceptible. By pinpointing these regions, they can strategically craft perturbations to trigger misclassifications. Attackers consistently exploit the identified vulnerable regions, such as those positions with larger mean alignment distance values or those with high index change frequency. The detection accuracy would remain high, around 100%. The converse is also true, attempting to craft successful adversarial examples solely within the robust regions – those with smaller mean alignment distance values or low index change frequency – poses significant challenges. The inherent tug-of-war between vulnerability and detection accuracy arises from the foundational property of consistent conceptual alignment. Given that these concepts encapsulate the quintessential patterns and nuances of the dataset, any misalignment due to perturbations will inevitably disrupt this harmony.

A2: Detector as a Discriminator. Another conceivable attack strategy would involve using the detector itself as a discriminator to guide adversarial perturbation crafting, a setting akin to Generative Adversarial Networks (GANs). This approach proved to be extremely challenging in successfully generating adversarial samples that could deceive both the classifier and the detector simultaneously; the success rate was below 1% (e.g., 8 out of 10,000), with substantial computational resources consumed. Secondly, adaptive samples that did bypass the system were those that are generally difficult to recognize in nature from the testing dataset, as illustrated in Figure 16 of the Appendix.

To counteract detector-adaptive attacks, a multi-pronged strategy can be implemented. By training a series of detectors and a set threshold for adversarial scores (for example, detector confidence). If an input's computed adversarial score overshoots this threshold, the final prediction for that input should be derived from a weighted majority voting mechanism. The ensemble can be diversified in multiple ways: using

autoencoder models derived from various checkpoints during DNN-GP training, combining autoencoders from differing architectures, or combining detection techniques.

6.3 Data Drift Detection

We expanded our evaluation to encompass both the MNIST-C and CIFAR-10-C datasets, renowned for their corrupted and perturbed samples. These samples undergo various alterations including noise, blur, and changes in brightness, leading to a systematic decline in the performance of the model. We interpret these samples as data drifts deviating from the original dataset. To detect these drift samples, we trained a One-Class SVM model using the original samples alone. Our detection accuracy reached 100% for both MNIST-C and CIFAR-10-C, maintaining a false positive rate below 5%.

7 Takeaways and Limitations

We revisit and summarize the key takeaways from DNN-GP and present possible limitations.

• **Latent Concept Understanding.** The aligned latent concept distance and index matrices effectively capture the essence of the original dataset and malicious patterns introduced by faults, enabling clear differentiation.

• **Specific Concept Patterns.** Distinct concepts were observed in each of original, adversarial, and semantic data. Recognizing these offers a deeper understanding of attack techniques.

• **Spatial Pattern Abstraction.** The position-aware metrics of the index matrix, successfully discerns high-level abstraction of spatial patterns in the pixel space. This provides a more intuitive understanding of how faulty samples are perceived and the image is spatially processed.

• **Impact of Positions.** The position-aware distance and spectral metrics reveal that specific positions in the latent concept matrix exhibit greater deviations when faults were present.

• **Spectral Feature Insights.** The spectral perspective of concept distances distinguishes between structural (low-frequency components) and intricate (high-frequency components) attributes. Perturbations primarily influence high-frequency components, whose subtle discrepancies present in the raw distance domain are amplified by spectral analysis.

• **Pixel Space Structural Patterns.** DNN-GP's reconstruction forcefully aligns non-structural pixel perturbations to expressed structural patterns, offering a window into the model's conceptual faults.

• **Limitations.** While most diagnostic results are based on the entire testing set, class-specific patterns are not distinctly evident. Although hypothesis testing validates the distinction between malicious and benign entities through alignment distance matrix metrics, the specific patterns contributing to this differentiation remain insufficiently explored. For instance, simple statistical measures such as the mean and variance of sample-wise distances prove inadequate. There is a clear need to employ more sophisticated statistical methodologies to achieve a deeper understanding. Nonetheless, DNN-GP is a versatile open tool, and with more refined settings, such as class-wise, or sample-wise analysis, additional patterns potentially with causality could possibly emerge. We have shown basic sample investigations.

8 Conclusion and Future Work

By proposing DNN-GP, we have addressed research gaps within abstraction and integration when interpreting DNN faults. By eclipsing traditional instance-level interpretations and delving deeper into the conceptual realm, our research paves the way for more resilient, transparent, and trustworthy deep learning systems. This study highlights the importance of the latent concept space in understanding and diagnosing the vulnerabilities found within DNNs. Our contributions include providing a comprehensive dataset for evaluating model vulnerability, creating an integrated diagnostic tool, establishing a new research direction by mapping probing samples into a conceptual space, and providing a valuable step in enhancing model resilience.

Moving forward, the latent concept space can be used for improved adversarial training to develop DNNs more robust to adversarial attacks and less prone to errors induced by data drifts. Further, one can incorporate larger models like transformers into the codebook training to potentially unlock richer conceptual spaces to enhance the quality of the conceptual fault diagnosis, detection and mitigation. Our evaluation of DNN-GP has been performed on image classification, but DNN-GP not confined to this domain. There are promising avenues to expand the application of DNN-GP to malware detection or large language models.

Finally, we publicly release datasets, code, diagnostic results, and detection results at https://github.com/TASI-LAB/DNN-GP.

Acknowledgments

This work is partly supported by Australian Research Council (ARC) DP240103068. Minhui Xue is supported by CSIRO – National Science Foundation (US) AI Research Collaboration Program. Shuo Wang and Minhui Xue are corresponding authors. We thank Haonan Zhong for the initial discussion of this paper.

References

[1] ABDI, H., AND WILLIAMS, L. J. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics 2*, 4 (2010), 433–459.

[2] AKHTAR, N., AND MIAN, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access 6* (2018), 14410–14430.

[3] AKULA, A., WANG, S., AND ZHU, S.-C. Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), vol. 34.

[4] ALDAHDOOH, A., HAMIDOUCHE, W., FEZZA, S. A., AND DEFORGES, O. Adversarial example detection for DNN models: A review and experimental comparison. *Artificial Intelligence Review* (2022).

[5] ANDRIUSHCHENKO, M., CROCE, F., FLAMMARION, N., AND HEIN, M. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision* (2020), Springer, pp. 484–501.

[6] ARJOVSKY, M., CHINTALA, S., AND BOTTOU, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning* (2017), PMLR.

[7] BAU, D., ZHU, J.-Y., STROBELT, H., LAPEDRIZA, A., ZHOU, B., AND TORRALBA, A. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences 117*, 48 (2020), 30071–30078.

[8] BEAM, A. L., AND KOHANE, I. S. Big data and machine learning in health care. *Jama 319*, 13 (2018), 1317–1318.

[9] BREUNIG, M. M., KRIEGEL, H.-P., NG, R. T., AND SANDER, J. LOF: Identifying density-based local outliers. In *2000 ACM SIGMOD International Conference on Management of Data* (2000), pp. 93–104.

[10] BRIGHAM, E. O. *The fast Fourier transform and its applications*. Prentice-Hall, Inc., 1988.

[11] BROWN, T. B., MANÉ, D., ROY, A., ABADI, M., AND GILMER, J. Adversarial patch. *arXiv preprint arXiv:1712.09665* (2017).

[12] CARLINI, N., ATHALYE, A., PAPERNOT, N., BRENDEL, W., RAUBER, J., TSIPRAS, D., GOODFELLOW, I., MADRY, A., AND KURAKIN, A. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705* (2019).

[13] CARLINI, N., AND WAGNER, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* (2017), pp. 3–14.

[14] CARLINI, N., AND WAGNER, D. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)* (2017), IEEE.

[15] CHEN, L., CHEN, J., HAJIMIRSADEGHI, H., AND MORI, G. Adapting grad-cam for embedding networks. In *IEEE/CVF Winter Conference on Applications of Computer Vision* (2020), pp. 2794–2803.

[16] DENG, L., YU, D., ET AL. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing 7*, 3–4 (2014), 197–387.

[17] ENGSTROM, L., TRAN, B., TSIPRAS, D., SCHMIDT, L., AND MADRY, A. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning* (2019), PMLR.

[18] GAMA, J., ŽLIOBAITĖ, I., BIFET, A., PECHENIZKIY, M., AND BOUCHACHIA, A. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR) 46*, 4 (2014), 1–37.

[19] GEIRHOS, R., RUBISCH, P., MICHAELIS, C., BETHGE, M., WICHMANN, F. A., AND BRENDEL, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231* (2018).

[20] GOODFELLOW, I. J., SHLENS, J., AND SZEGEDY, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[21] GRAPOV, D., FAHRMANN, J., WANICHTHANARAK, K., AND KHOOMRUNG, S. Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. *Omics: A journal of integrative biology 22*, 10 (2018), 630–636.

[22] GULRAJANI, I., AHMED, F., ARJOVSKY, M., DUMOULIN, V., AND COURVILLE, A. C. Improved training of wasserstein gans. *Advances in Neural Information Processing Systems 30* (2017).

[23] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.

[24] HENDRYCKS, D., AND DIETTERICH, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261* (2019).

[25] HOSSEINI, H., AND POOVENDRAN, R. Semantic adversarial examples, 2018.

[26] HUANG, G., LIU, Z., VAN DER MAATEN, L., AND WEINBERGER, K. Q. Densely connected convolutional

networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017).

[27] JOSHI, A., MUKHERJEE, A., SARKAR, S., AND HEGDE, C. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *IEEE/CVF International Conference on Computer Vision* (2019), pp. 4773–4783.

[28] KIM, B., WATTENBERG, M., GILMER, J., CAI, C., WEXLER, J., VIEGAS, F., ET AL. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning* (2018), PMLR.

[29] KOTYAN, S., AND VARGAS, D. V. Adversarial robustness assessment: Why in evaluation both l_0 and $l$inf attacks are necessary. *PloS One 17*, 4 (2022), e0265723.

[30] KRIZHEVSKY, A., HINTON, G., ET AL. Learning multiple layers of features from tiny images.

[31] LECUN, Y., BOTTOU, L., BENGIO, Y., AND HAFFNER, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324.

[32] LI, L., LV, Y., AND WANG, F.-Y. Traffic signal timing via deep reinforcement learning. *IEEE/CAA Journal of Automatica Sinica 3*, 3 (2016), 247–254.

[33] LI, L., XIE, T., AND LI, B. Sok: Certified robustness for deep neural networks. In *2023 IEEE Symposium on Security and Privacy (SP)* (2023), IEEE.

[34] LIU, Z., LUO, P., WANG, X., AND TANG, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 3730–3738.

[35] LU, J., LIU, A., DONG, F., GU, F., GAMA, J., AND ZHANG, G. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering 31*, 12 (2018), 2346–2363.

[36] LUNDBERG, S. M., AND LEE, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30* (2017).

[37] MA, S., LIU, Y., TAO, G., LEE, W.-C., AND ZHANG, X. Nic: Detecting adversarial samples with neural network invariant checking. In *26th Annual Network And Distributed System Security Symposium (NDSS 2019)* (2019), Internet Soc.

[38] MA, X., LI, B., WANG, Y., ERFANI, S. M., WIJEWICKREMA, S., SCHOENEBECK, G., SONG, D., HOULE, M. E., AND BAILEY, J. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613* (2018).

[39] MADRY, A., MAKELOV, A., SCHMIDT, L., TSIPRAS, D., AND VLADU, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[40] MASSEY JR, F. J. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association 46*, 253 (1951), 68–78.

[41] MENG, D., AND CHEN, H. Magnet: a two-pronged defense against adversarial examples. In *2017 ACM SIGSAC Conference on Computer and Communications Security* (2017), pp. 135–147.

[42] MOOSAVI-DEZFOOLI, S.-M., FAWZI, A., AND FROSSARD, P. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 2574–2582.

[43] MU, N., AND GILMER, J. Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337* (2019).

[44] NGUYEN, H., KIEU, L.-M., WEN, T., AND CAI, C. Deep learning methods in transportation domain: a review. *IET Intelligent Transport Systems 12*, 9 (2018), 998–1004.

[45] PAPERNOT, N., MCDANIEL, P., JHA, S., FREDRIKSON, M., CELIK, Z. B., AND SWAMI, A. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)* (2016), IEEE, pp. 372–387.

[46] PAPERNOT, N., MCDANIEL, P., SINHA, A., AND WELLMAN, M. P. Sok: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)* (2018), IEEE.

[47] RAZAVI, A., VAN DEN OORD, A., AND VINYALS, O. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems* (2019), pp. 14866–14876.

[48] ROUSSEEUW, P. J., AND DRIESSEN, K. V. A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41*, 3 (1999), 212–223.

[49] SCHÖLKOPF, B., PLATT, J. C., SHAWE-TAYLOR, J., SMOLA, A. J., AND WILLIAMSON, R. C. Estimating the support of a high-dimensional distribution. *Neural Computation 13*, 7 (2001), 1443–1471.

[50] SELVARAJU, R. R., COGSWELL, M., DAS, A., VEDANTAM, R., PARIKH, D., AND BATRA, D. Grad-cam:

Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 618–626.

[51] SHALEV-SHWARTZ, S., AND BEN-DAVID, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[52] SOTGIU, A., DEMONTIS, A., MELIS, M., BIGGIO, B., FUMERA, G., FENG, X., AND ROLI, F. Deep neural rejection against adversarial examples. *EURASIP Journal on Information Security* (2020).

[53] TOLSTIKHIN, I., BOUSQUET, O., GELLY, S., AND SCHOELKOPF, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558* (2017).

[54] TSCHANNEN, M., BACHEM, O., AND LUCIC, M. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069* (2018).

[55] VAN DEN OORD, A., VINYALS, O., ET AL. Neural discrete representation learning. In *Advances in Neural Information Processing Systems* (2017).

Appendix

In this section, we provide additional background for adversarial attacks and defenses. We also present a evaluation on the importance of cluster-based alignment, details on datasets and models, and additional results. Unfortunately, due to space limitations, additional details are available in the full version of the Appendix presented in the GitHub repository https://github.com/TASI-LAB/DNN-GP.

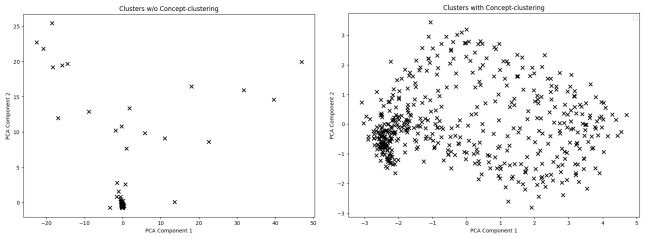
A Probing Approaches and Settings

Vulnerabilities derived from adversarial and semantic attacks involves the manipulation of input data intended for machine learning models, that induce errors [33]. One common example is the realm of image classification, where small, carefully crafted changes to an image can cause a state-of-the-art model to misclassify [20]. This phenomenon raises serious concerns about the robustness and reliability of machine learning models, particularly in critical applications such as autonomous driving or cybersecurity [2].

Adversarial Attacks. Full description of adversarial attacks of FGSM, PGD, JSMA, DeepFool, CW, and the Pixel Attack are available in the full version of the Appendix in GitHub.

Semantic Attacks. Semantic attacks [27] on DNNs are a relatively new area of research. These attacks exploit the model's inability to understand the semantic meaning of the input data. In the context of image recognition, a semantic attack might involve manipulating transformation $T(\cdot)$, such as the rotation or lighting of an image, to fool the model into misclassification, i.e., $f(T(\boldsymbol{x})) \neq f(\boldsymbol{x})$ [11, 17]. Unlike traditional adversarial attacks that add imperceptible noise to the input, semantic attacks modify the image in a way that is perceptible but semantically plausible to humans. The detailed description of semantic attacks contrast change, rotation, and color-shift against image classification are provided in the full version of the Appendix in the GitHub repository.

Data Drift: Instance Corruptions. Another typical semantic attack against DNNs in the image domain is the introduction of corruptions that preserve the semantic content of the image but significantly degrade the performance of state-of-the-art computer vision models. This approach is demonstrated in MNIST-C dataset, a comprehensive suite of 15 corruptions applied to the MNIST test set, for benchmarking out-of-distribution robustness in computer vision [43]. The corruptions are model-agnostic and do not seek worst-case performance. Instead, they are designed to be broad and diverse, capturing multiple failure modes of modern models.

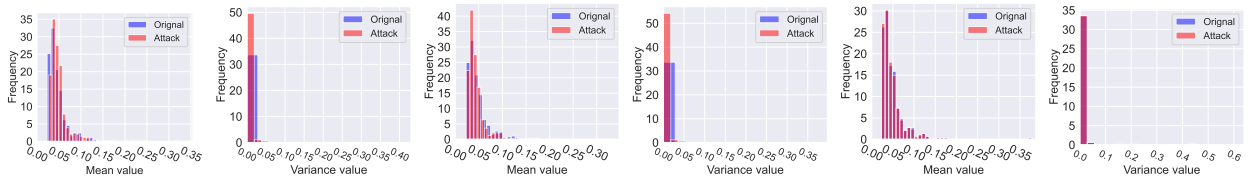


(a) W/o clustering alignment. (b) With clustering alignment.

Figure 11: Visualization of concept items in the codebook with and without clustering-based alignment.

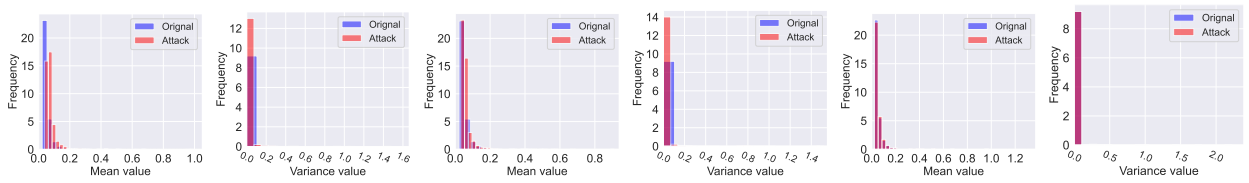
B Clustering-based Alignment

We underscore the importance of employing a clustering-based alignment strategy, with the MNIST dataset as an illustrative example as its latent space is more straightforward to interpret. Figure 11 demonstrates that without the clustering-based alignment, a majority of codebook concept items will cluster closely, rendering them unused in the alignment for both original and adversarial samples. Only the remaining sparse concepts are actively utilized in alignment. This results in original and adversarial samples index matrices being highly similar, complicating the task of concept-based diagnosis. In contrast, after implementing cluster alignment, the concept items are more evenly distributed. Moreover, there is a pronounced difference in the concepts used in the alignment for original and adversarial samples, both in terms of index and alignment distance. This enhances the identification of unique concepts that are specifically employed during adversarial sample alignment, as compared to the original.



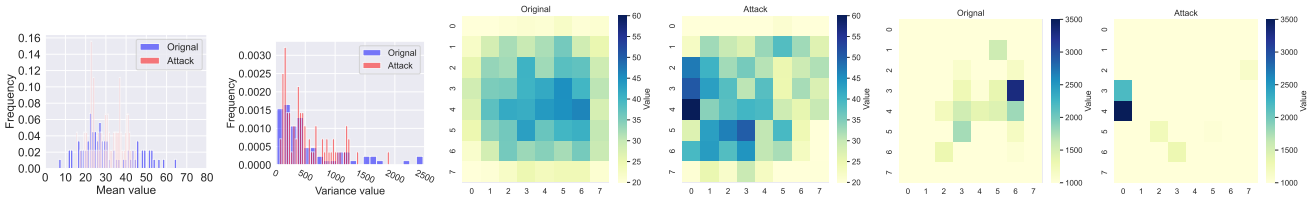
(a) Mean (FGSM-32) (b) Var. (FGSM-32) (c) Mean (PGD-32) (d) Var. (PGD-32) (e) Mean (CW, inf) (f) Var. (CW, inf)

Figure 12: Sample-wise distance distributions of original and adversarial samples on *DenseNet-121* trained on CelebA.



(a) Mean (FGSM-32) (b) Var. (FGSM-32) (c) Mean (PGD-32) (d) Var. (PGD-32) (e) Mean (CW, inf) (f) Var. (CW, inf)

Figure 13: Sample-wise distance distributions of original and adversarial samples on *ResNet-18* trained on CelebA.



(a) Mean (b) Variance (c) Position-aware mean (d) Position-aware variance

Figure 14: Sample-aware and position-aware distance distributions of original and drift samples on CIFAR-10.

C Datasets and Models

In the experiments, we use benchmark image datasets of MNIST [31], CIFAR-10 [30], and CelebA [34]. We use a 7-layer ordinary CNN model for MNIST, a 9-layer ordinary CNN model for CIFAR-10, and a 5-layer ordinary CNN model for CelebA. Due to page limits, we provide detailed description of datasets and model architecture in the GitHub repository at https://github.com/TASI-LAB/DNN-GP.

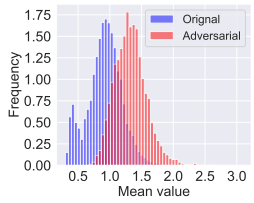
D Extended Diagnosis Results

D.1 Additional Results on MNIST and CelebA

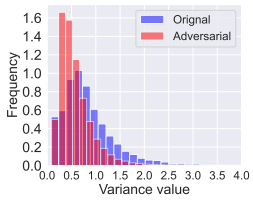
In this section, we provide typical diagnostic results on MNIST, including KS test results in Table 3 and a diagnose case of FGSM when the target model is a 9-layer ordinary CNN model in Figure 15. In addition, to show DNN-GP's model-agnostic property, we provide additional diagnostic results on CelebA when the target model is ResNet-18 or DenseNet-121 attacked by three adversarial attacks of FGSM, PGD, and CW in Figure 13 and Figure 12. All these results confirm the findings in CIFAR-10 in the main text. Complete results for all datasets across various faulty settings can be accessed within our repository at GitHub.

Table 3: KS test diagnosis results under multiple adversarial attacks for the MNIST

Attack Setting		D-value	p-value
Attack Method	Noise Parameter		
FGSM	8/255	0.017	0.028
	16/255	0.014	1.712×10^{-4}
	32/255	0.009	1.489×10^{-4}
	64/255	0.020	1.213×10^{-6}
	80/255	0.048	5.858×10^{-19}
PGD-Linf	8/255	0.013	1.329×10^{-33}
	16/255	0.043	≈0
	32/255	0.124	≈0
	64/255	0.347	≈0
	80/255	0.502	≈0
CW-Linf	-	0.017	2.6×10^{-8}
Pixel	3	0.024	9.783×10^{-13}
DeepFool	-	0.05	≈0
JSMA	-	0.143	≈0



(a) Mean (FGSM, 64)



(b) Variance (FGSM, 64)

Figure 15: Sample-wise distance distributions of original and adversarial samples with FGSM on MNIST.

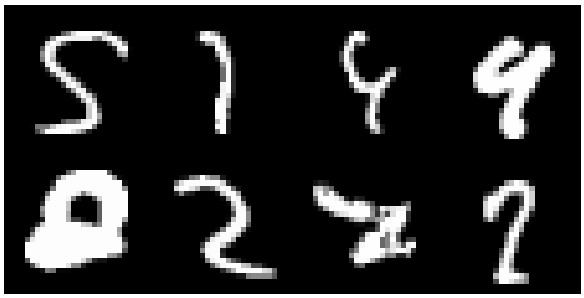


Figure 16: Hard samples recovered from false positive set.

D.2 Diagnosis on Data Drift

In this section, we present the diagnostic results for MNIST-C and CIFAR10-C, as shown in Figure 14. These results align with our findings from the semantic attacks, as they are similar in transformations but without verifying if each sample results in a misclassification for MNIST-C and CIFAR10-C. Complete results for all datasets across various faulty settings are provided in GitHub.

D.3 Reconstruction of Samples

In this section, we provide reconstruction visualizations of original and adversarial samples in Figure 17 to understand how the reconstructed versions of the attacked images can shed light on the meaning of latent concept the attacker attempts to exploit. The complete reconstruction visualizations for MNIST-C and semantic attack samples in CIFAR-10 and CelebA are provided in the GitHub link. We also report the classification performance after reconstructions in Table 4.

D.4 Additional Detection Results

We also study the detection performances of DNN-GP on MNIST and CelebA. We find that DNN-GP works well on detecting faulty samples on MNIST and CelebA, aligning with our findings on CIFAR-10 in Table 2. Detailed tables are provided in the GitHub.

Table 4: Prediction accuracy on faulty MNIST samples before and after reconstruction by DNN-GP. As we can see, DNN-GP has a certain ability to purify faulty samples.

Attack Setting		Before	After	Original	Ori. Reconstruct
Attack	Noise				
FGSM	32/255	0.00	0.84	0.98	0.96
	64/255	0.00	0.77	0.98	0.96
	80/255	0.00	0.66	0.98	0.96
PGD-Linf	32/255	0.00	0.95	0.98	0.96
	64/255	0.00	0.85	0.98	0.96
	80/255	0.00	0.71	0.98	0.96
CW-Linf	-	0.00	0.70	0.98	0.96
Pixel	-	0.00	0.75	0.98	0.96
DeepFool	-	0.00	0.91	0.98	0.96
JSMA	-	0.00	0.68	0.98	0.96

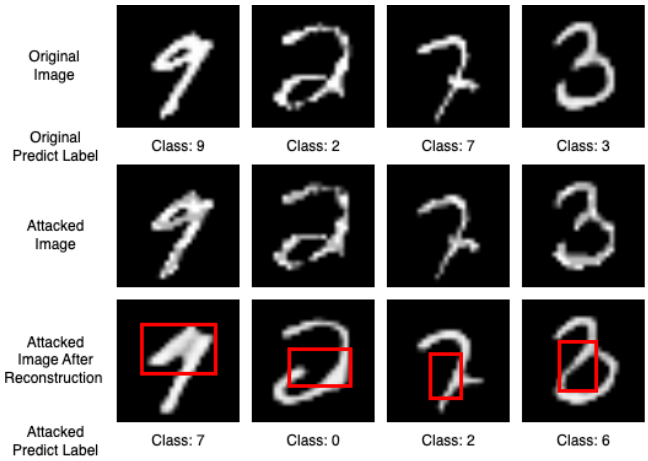


Figure 17: Reconstruction of Adversarial Examples from MNIST using FGSM-32/255. The attack on the first image fills in blank spaces of the number 9 to forge a 7. The tail of number 2 in the 2nd image is weakened, leading to its reconstruction as a 0. The third attack shifts the upper stroke of the number 7 to the right, connecting it at the tail, resulting in a 2 classification. The fourth merges the bottom half of the number 3, into a circular shape resembling the number 6. The adversarial samples show no damage the original image's structure and content, with noise spread across the entire number, the reconstruction revealing the basic attack patterns (i.e. in our MNIST examples, by altering small strokes, a transition to another number with a similar structure), and also within remaining stroke areas, reconstruction removes impacts of adversarial noise.