Report on Stress Testing AI Models with Casual Subversion Attack

Sindhu Pasupuleti

May 2025

0.1 Problem Statement

As AI systems become integral to critical domains such as autonomous vehicles, medical diagnosis and content moderation, understanding and securing their behavior in unpredictable environments is more urgent than ever. While current deep learning models achieve high accuracy on standard benchmarks, they remain brittle when exposed to out-of-distribution (OOD) inputs or semantically adversarial examples. These vulnerabilities typically lie concealed behind normal tests but have the ability to result in catastrophic failures upon real deployment. Current robustness protocols largely focus on low level noise or simple perturbations, overlooking the deeper semantic and causal structures central to human perception.

Our research introduces Causal Concept Subversion (CCS), a novel method for probing the conceptual robustness of vision models by targeting their reliance on causally critical features. Unlike traditional adversarial approaches that rely on pixel-level noise or synthetic distortions, CCS strategically alters semantically meaningful content within an image. We first use Grad-CAM to identify regions most influential to a model's prediction. These regions are then replaced with patches that are perceptually similar (using LPIPS) but semantically conflicting (using CLIP embeddings), ensuring the perturbations are visually coherent yet conceptually disruptive.

To evaluate the model's sensitivity to such interventions, we propose the Conceptual Fragility Index (CFI), a metric that captures prediction volatility across multiple adversarial variants. CCS preserves overall image structure while subtly misleading the model's internal reasoning, exposing vulnerabilities in its causal understanding. This method offers a targeted, interpretable approach to assessing robustness in open-world, semantically complex scenarios where traditional in-distribution assumptions no longer hold.

0.2 Methodology

0.2.1 Datasets

This study employs two publicly available data sets: ImageNet-256 and the Intel Image Classification data set, collected from Kaggle.

The ImageNet-256 dataset is the downsampled variant of the original ImageNet dataset and consists of images from 256 object classes. There are between 1,000 and 5,000 RGB images per class, and a total of one million images. The images are standardized to a resolution of 256×256 pixels.

The Intel Image Classification dataset has 25,000 natural scene images categorized into six classes: buildings, forest, glacier, mountain, sea and street. Each class contains 3000 - 5000 images and are resized to 150×150 pixels. This dataset is often used for benchmarking scene classification models and algorithms.

0.2.2 Causal Concept Subversion (CCS)

A framework called **Causal Concept Subversion (CCS)** is proposed to evaluate the fragility of a classifier by selectively perturbing causal regions identified via Grad-CAM. The goal is to introduce carefully selected out-of-distribution (OOD) patches into the causally important areas of the input, maximizing both low-level texture difference and high-level semantic drift, and to measure how unstable the model's predictions become under these controlled perturbations.

Experimental Procedure

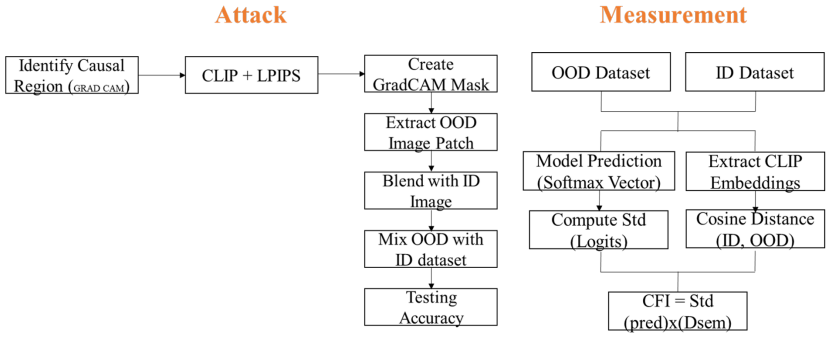

Figure 1: Methodology for CCS

The procedure for the attack and measurement is as follows:

1. Load ResNet-18 / EfficientNet-B0, CLIP ViT-B/32, and LPIPS AlexNet models.
2. Normalize ID dataset images to ImageNet statistics and preprocess OOD images appropriately for LPIPS and CLIP feature extraction.
3. For each ID image:
   - Generate GradCAM heatmaps and threshold to create causal masks.
   - Search for OOD patches maximizing the sum of LPIPS distance and CLIP distance.
   - Blend the selected OOD patch into the causal regions with blending factor α, and classifies the perturbed images using model.
   - Compute semantic shift between original and perturbed images.
   - Aggregate predictions and semantic shifts to compute CFI.
4. Analyze the results. A higher CFI indicates greater model fragility under causal concept subversion.

Preprocessing

Out-of-Distribution (OOD) Patches: OOD images serve as perturbation patch sources. Each OOD image is resized to 224×224 pixels to match the model input size. For the computation of the LPIPS texture distance, these images are further normalized to the interval $[-1, 1]$. Images are fed into CLIP's preprocessing pipeline (resize, center crop as well as normalization) for CLIP semantic feature extraction.

Feature Extraction

Texture Feature (LPIPS Distance): A pre-trained LPIPS model with an AlexNet backbone is used to compute perceptual similarity between the original image and OOD candidate patches. A larger LPIPS distance reflects a more significant low level texture difference useful in distinguishing visually different perturbation candidates.

Semantic Feature (CLIP Embeddings): CLIP ViT-B/32 is used to achieve semantic embeddings of both the original images and OOD patches. Semantic dissimilarity between them is calculated as the cosine distance between their corresponding embeddings.

Models and Parameters

- **ResNet-18 / EfficientNet-B0** (pretrained on ImageNet): used for predictions and Grad-CAM extraction; both models are frozen (`model.eval()` mode).
- **CLIP ViT-B/32**: used for semantic feature extraction; used without fine-tuning.
- **LPIPS Model** (AlexNet backbone): used for perceptual texture distance measurement.
- **Grad-CAM Setup:** target layers are `layer4[-1]` for ResNet-18 and `model.features[-1]` for EfficientNet-B0.

Hyperparameters: Number of perturbations per image: $N_{\text{variants}} = 5$ and Blending factor for patch and causal region: $\alpha = 0.75$.

Formulas Used

Texture Dissimilarity (LPIPS Distance):

$$\text{LPIPS}(x, y) = \sum_{l} \frac{1}{H_l W_l} \|F_l(x) - F_l(y)\|_2^2$$

where $F_l(\cdot)$ are feature activations at layer l of a network.

Semantic Dissimilarity (CLIP Distance):

$$\text{CLIP\_Dist}(x, y) = 1 - \cos(\theta) = 1 - \frac{x \cdot y}{\|x\|\|y\|}$$

where x and y are CLIP feature embeddings.

Total Patch Selection Score:

$$\text{Total Score} = \text{LPIPS Distance} + \text{CLIP Distance}$$

The patch with the highest total score is selected.

Conceptual Fragility Index (CFI):

$$\text{CFI} = \sigma(\text{Predictions}) \times \mu(\text{Semantic Shifts})$$

where:

- σ(Predictions) is the mean standard deviation of prediction probabilities across all perturbed variants,
- μ(Semantic Shifts) is the average semantic shift (cosine distance) across variants.

0.2.3 Results of Causal Concept Subversion

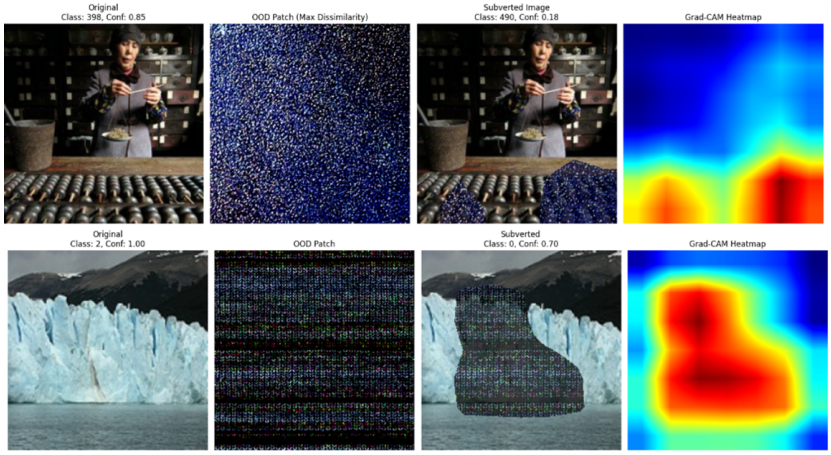

Figure 2: Output of ResNet18 on ImageNet and Intel Classification Dataset For the majority of cases, combining an out-of-distribution (OOD) patch with an in-distribution image effectively decreases the model's prediction confidence demonstrating the intended impact of our adversarial method. For instance, in the figure 11, we show a person using an abacus. ResNet-18 initially classifies the image with high confidence (0.85), but once an OOD patch is blended into the causal region, confidence drops sharply to 0.18, showing successful disruption of the model's internal reasoning. And same with the image from Intel Classification.

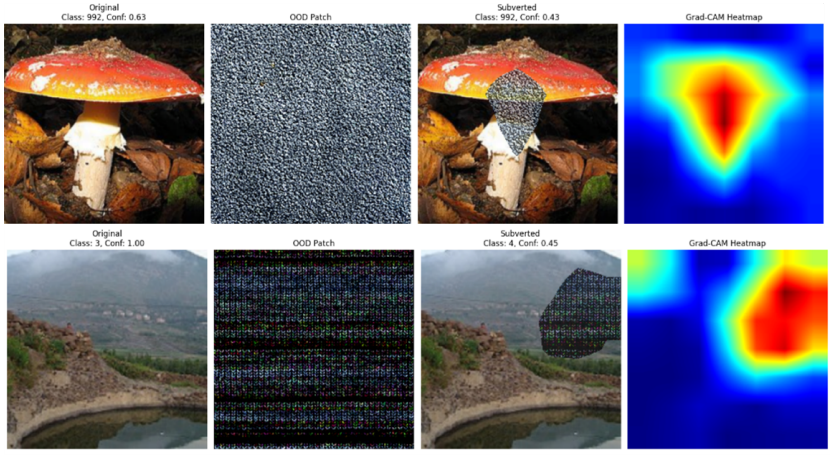

Figure 3: Output of EfficientNet-B0 on ImageNet Dataset

In Figures 12, we observe the same tendencies but EfficientNet-B0 is stronger. Its predictions are less confident overall but it also predicts classes correctly, and the entropy is greater even after patch insertion, which demonstrates a better calibrated uncertainty response. This suggests EfficientNet is better at signaling ambiguity when exposed to OOD content, making it more suitable for safety-critical applications.

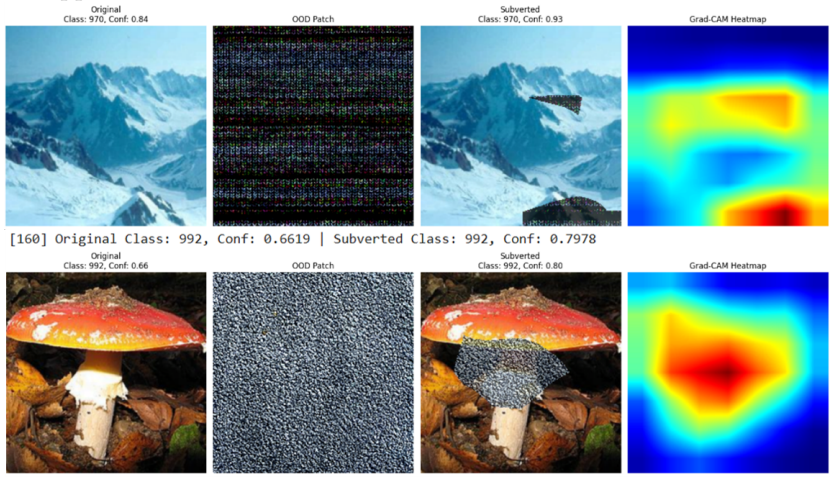

Figure 4: Output of ResNet18 on ImageNet Dataset and EfficientNet-B0 on Intel Classification Dataset

However, there are certain cases that are counterintuitive. In the second image of Figure 13 with mushroom and landscape introducing semantically irrelevant patches unexpectedly increased model confidence. Such unexpected behavior reflects a vulnerability: models like ResNet18 and EfficientNet-B0 can spuriously become confident when patch textures incidentally match learned local features.

This highlights a broader issue that such models heavily rely on local texture information rather than global scene understanding. Even absurd patches can support existing evidence if their surface statistics align with class priors. While EfficientNet is generally more conservative, it also shows overconfidence under such conditions pointing towards the vulnerability of today's CNNs to localized, texture based OOD attacks and the need for models with stronger global reasoning capabilities.

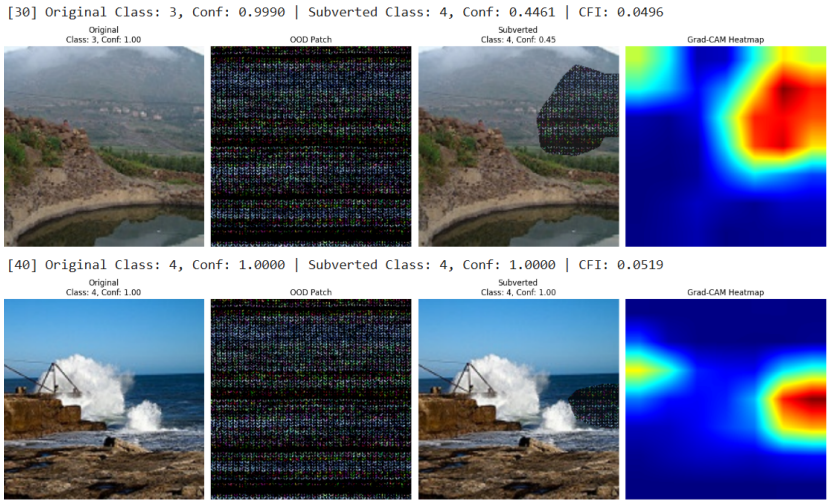

Figure 5: CFI Score for ResNet18 on Intel Classification Dataset

The Causal Fragility Index or CFI captures how quickly the prediction of a model changes upon adding a small, localized patch to an image. Low CFI in the case of the examples (0.0496 and 0.0519) suggests the decision of the model is quite stable and not dependent much on the patch. This makes CFI a useful metric for assessing the robustness and causal reliability of visual models under perturbations. This shows that ResNet18 is quite stable from casual perturbations in this dataset.

0.3 Discussion

The primary objective of this work was to evaluate the robustness of convolutional neural networks, specifically ResNet-18 and EfficientNet-B0, against conceptually targeted adversarial interventions through our proposed Causal Concept Subversion (CCS) method. This approach aimed to probe the models' reliance on causally significant features by introducing localized, semantically conflicting yet perceptually plausible perturbations. Our working hypothesis was that despite their architectural sophistication, modern CNNs remain fragile when causal reasoning is subverted at critical input regions.

The results validated this hypothesis. CCS-induced perturbations significantly degraded both models' predictive performance, with ResNet-18 showing

greater vulnerability. Notably, ResNet-18 exhibited consistent overconfidence, maintaining high prediction probabilities even when causal regions were replaced with deceptive patches. In contrast, EfficientNet-B0 demonstrated a relatively better ability to express uncertainty, reflected in higher entropy and reduced confidence in its predictions on perturbed inputs. This supports the broader observation that architectural improvements like compound scaling in EfficientNet may indirectly benefit robustness, albeit modestly.

To quantify this degradation, we introduced the Conceptual Fragility Index (CFI), which measures the volatility of model predictions across multiple CCS variants. Higher CFI scores indicated a higher sensitivity to causal perturbations. EfficientNet-B0 consistently showed lower CFI scores than ResNet-18, suggesting more stable internal reasoning under conceptually adversarial conditions.

Despite these promising insights, the current work has limitations. First, the study focuses solely on image-based CNNs. This restricts generalizability to other model types, such as transformers or multimodal systems that might use richer contextual cues for inference. Second, the patch-based subversion pipeline, while effective in exposing vulnerabilities, relies on manual hyperparameter tuning and handcrafted patch selection. This limits the scalability and adaptability of the attack, especially in dynamic real-world environments. Furthermore, while LPIPS and CLIP help preserve visual plausibility and enforce semantic divergence, the introduction of synthetic patches may still lack the nuanced complexity of real-world distributional shifts.

Future work should extend CCS in three key directions. First, integrating generative models or reinforcement learning could automate patch generation and improve adaptability across domains. Second, deploying CCS across diverse tasks and modalities—such as medical imaging, autonomous driving, or cross-modal VQA systems—could test robustness in settings where causal reasoning is critical. Finally, incorporating human-in-the-loop evaluations to improve the realism and contextual relevance of perturbations may offer richer insights into the fragility of model reasoning. Longitudinal testing during model retraining cycles would further reveal how CCS-type perturbations influence learning dynamics and real-world resilience.

0.4 Conclusion

In this work, we introduced Causal Concept Subversion (CCS), a novel adversarial strategy designed to evaluate and stress-test the causal robustness of vision models by targeting semantically coherent yet causally disruptive regions in images. Our method identifies causally significant regions using Grad-CAM, then replaces these areas with visually plausible but semantically contradictory patches. These substitutions are guided by LPIPS and CLIP-based distances to ensure perceptual consistency while intentionally misleading the model's internal reasoning. To quantify the impact, we proposed the Conceptual Fragility Index (CFI), a new metric that captures the instability of model predictions

under causal perturbations.

We evaluated CCS on two widely used CNN architectures, ResNet-18 and EfficientNet-B0, and observed a significant decline in performance and prediction confidence, particularly for ResNet-18. While EfficientNet-B0 exhibited misclassifications, it displayed greater uncertainty on perturbed inputs, suggesting a relatively stronger calibration. The results highlight a critical vulnerability in modern vision models—their overreliance on local textures and insufficient causal reasoning. In some cases, models remained highly confident in incorrect predictions, revealing a misalignment between visual integrity and semantic understanding.

These findings underscore the need for robust model interpretability and reliability, especially in high-stakes settings like autonomous driving, medical diagnostics, and security surveillance, where causal reasoning is essential. Our work demonstrates that even perceptually subtle perturbations, when applied to causally critical regions, can destabilize model outputs and expose conceptual weaknesses.

Future work could explore automating CCS via generative adversarial techniques, extending it to multimodal or transformer-based models, and evaluating the method in more ecologically valid, real-world environments. Overall, this work contributes a valuable diagnostic tool for probing conceptual robustness and promoting the development of more resilient AI systems.