Report on Stress Testing AI Models with Semantic Confusion 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.

We introduce Semantic Confusion, a generative out-of-distribution (OOD) attack that crafts misleading yet visually plausible images to probe model robustness. This method begins by using a BART language model to generate semantically unusual or contradictory captions (e.g., "a zebra rug in a living room"). These captions are then passed to a diffusion-based image synthesis model to produce realistic, yet semantically conflicting visuals. The goal is to expose generalization failures in image classifiers such as ResNet-18 when faced with conceptually distorted but natural-looking inputs.

To evaluate how well a model maintains semantic coherence under such attacks, we propose the Perceptual Concept Shift (PCS) metric. PCS quantifies the degree of semantic misalignment by comparing Grad-CAM attention heatmaps, region-level captions generated by BLIP, and CLIP-based image-caption similarity scores. By triangulating these three perspectives, PCS provides a structured and interpretable measure of how conceptually grounded a model's predictions remain in the presence of confusing, hybrid content. Together, Semantic Confusion and PCS offer a powerful framework for stress-testing vision models in open-world scenarios where compositional and conceptual deviations frequently arise.

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 Semantic Confusion (SC)

Semantic Confusion (SC) is an attack framework for measuring model robustness against semantically anomalous but visually plausible inputs. It evaluates a model's susceptibility to subtle contextual violations by generating such instances and comparing performance on EfficientNet-B0 and ResNet18.

Experimental Procedure

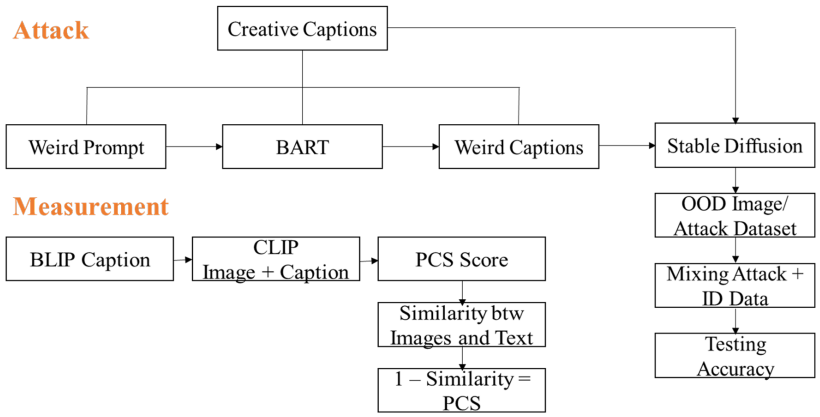

Figure 1: Methodology for SC

The Semantic Confusion attack and measurement procedure involves the following steps:

- Load ResNet-18, EfficientNet-B0, BLIP, and CLIP ViT-B/32 models.
- Normalize dataset images to match dataset statistics.
- Use BART to generate semantically unusual base prompts and stylized variants.
- For each generated prompt:
  - Generate an image using Stable Diffusion v1.4.
  - Classify the image using ResNet-18 and EfficientNet-B0 to get predictions and softmax probabilities.
  - Generate a GradCAM heatmap and extract the salient causal region.
  - Use BLIP to caption the GradCAM region.
  - Compute PCS by measuring cosine similarity between the BLIP caption embedding and the CLIP image embedding.

  - Measure softmax entropy and cosine distance from cached in-distribution (ID) embeddings.

- High PCS (> 0.3), entropy, and cosine distance indicate semantic confusion and possible OOD behavior.

Preprocessing

In-Distribution (ID) Datasets: The ID datasets used are the ImageNet-256 dataset and the Intel Image Classification dataset. All input images are resized to 256×256 pixels and center-cropped to 224×224 before being normalized using the dataset's statistics:

- Mean: [0.485, 0.456, 0.406]
- Std: [0.229, 0.224, 0.225]

The preprocessing for In-distribution dataset is same for all three methods. For visualization, images are "unnormalized" back to the pixel range $[0, 1]$.

Out-of-Distribution (OOD) Image Generation: A set of atypical or incongruent prompts (e.g., "a zebra rug in a living room") is fed into the `facebook/bart-large-cnn` model to generate variants, with decoding parameters: max_length = 20, temperature = 1.0, top_k = 50, num_beams = 5. These prompts are rendered into images using `Stable Diffusion v1.4` via the HuggingFace Diffusers pipeline, saved at 512×512 resolution in PNG format, and resized/normalized to match the ID input pipeline.

Feature Extraction

Image Features: For each sample (ID and generated), we extract a 512-dimensional feature vector $\mathbf{f}_{\text{img}}$ from the penultimate layer of ResNet-18 and EfficientNet-B0:

$$\mathbf{f}_{\text{img}} = \text{Backbone}_{\text{features}}(x), \quad \mathbf{f}_{\text{img}} \in \mathbb{R}^{512}$$

A reference pool of 512 randomly sampled ImageNet features is cached to estimate the distribution of in-domain feature embeddings for cosine distance calculations.

Region Features (GradCAM): GradCAM localizes semantically meaningful regions of the image for the predicted class. For ResNet-18, feature activations are extracted from `model.layer4[-1]` and for EfficientNet-B0 from `model.features[-1]`. These regions are described and compared to full image semantics. We will be using the same GradCAM extraction method for all three attacks.

Textual Semantics (BLIP and CLIP): BLIP is used to caption the Grad-CAM area, generating a text summary of the causal part of the image. The captions, along with the original prompt or image caption, are embedded using `openai/clip-vit-base-patch32`. Semantic similarity is estimated by cosine similarity among both embeddings.

Models and Parameters

- **ResNet-18 and EfficientNet-B0:** Pretrained on ImageNet and used in `eval()` mode.
- **Stable Diffusion v1.4:** Used to synthesize realistic but semantically anomalous images.
- **BART (facebook/bart-large-cnn):** Rewrites prompts to increase linguistic creativity.
- **BLIP:** Used for region-level captioning.
- **CLIP ViT-B/32:** Used to compute text-image embedding similarity.
- **Grad-CAM:** Extracts causal regions from classifier predictions.

Formulas Used

Entropy (Classification Uncertainty):

$$H(p) = -\sum_{i=1}^{C} p_i \log(p_i + \epsilon), \quad \epsilon = 1 \times 10^{-10}$$

where p_i is the softmax probability for class i and C is the total number of classes.

Cosine Distance (Feature Deviation):

$$d_{\cos} = 1 - \frac{\mathbf{f}_g \cdot \mathbf{f}_{\text{id}}}{\|\mathbf{f}_g\| \|\mathbf{f}_{\text{id}}\|}$$

where $\mathbf{f}_g$ is the feature from the generated image and $\mathbf{f}_{\text{id}}$ is an in-distribution reference.

Perceptual Concept Shift (PCS):

$$\text{PCS} = 1 - \cos\left(\mathbf{v}_{\text{text}}, \mathbf{v}_{\text{image}}\right)$$

Here, $\mathbf{v}_{\text{text}}$ is the CLIP embedding of the BLIP-generated caption for the Grad-CAM region, and $\mathbf{v}_{\text{image}}$ is the CLIP embedding of the full image.

0.2.3 Results of Semantic Confusion Attack

This section presents a comparative analysis of the performance of ResNet18 and EfficientNet-B0 on semantically perturbed and out-of-distribution (OOD) images. We compare the predictions of the models in three representative scenarios to test their susceptibility to texture bias, their handling of semantically perturbed but in-distribution images and their response to OOD inputs. Results are visualized in Figures 1–3 and are interpreted alongside key metrics including confidence scores and prediction entropy.

Figure 2: ResNet18 and EfficientNet-B0 on ImageNet

In Figure 4, we show a zebra-print sofa that is designed to test the model's shape dependence over texture. ResNet18 classifies it as a zebra with very high confidence (¿98%) and low entropy indicating high confidence in its incorrect prediction as is characteristic of its widely documented texture bias. EfficientNet-B0 classifies it with low confidence (¡30%) and high entropy and indicates greater uncertainty. While it still misclassifies the image, its cautious behavior highlights its reduced susceptibility to confident errors, making it more suitable for detecting potential OOD inputs and improving safety in critical applications.

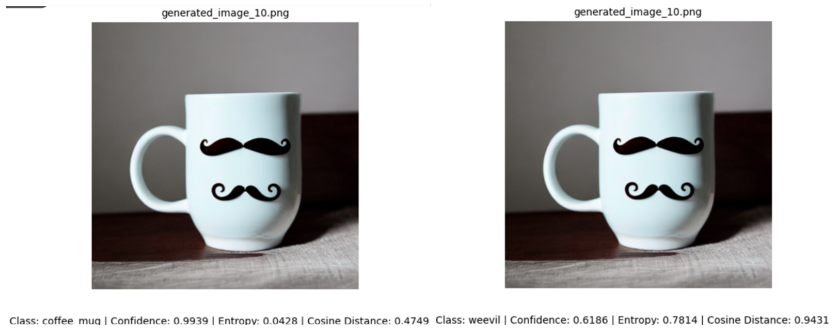

Figure 3: ResNet18 and EfficientNet-B0 on ImageNet

Figures 5 show a mustache-print coffee mug. ResNet18 correctly classifies mug with ¿99% confidence exhibiting robustness to minor semantic changes when shape is preserved. EfficientNet-B0 fails to recognize the object and provides a low confidence, high entropy prediction, being conservative but less accurate in in-distribution tasks.

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

In Figure 6, a forest painted van from the Intel Scene Classification dataset is misclassified as a glacier by both the models. ResNet18 again has low entropy and high confidence while EfficientNet-B0 once again predicts with high entropy and low confidence. This demonstrates the overconfidence of ResNet in OOD settings, while EfficientNet expresses uncertainty.

Overall ResNet18 is more accurate in-distribution but overconfident on ambiguous inputs. EfficientNet-B0 is more conservative and therefore better suited for applications where uncertainty estimation is important.

```
--------------------------------------------------
Image: /kaggle/input/testset/generated_image_17.png
Caption: a pizza with pepperoni and pepperoni on it
PCS Score: 0.30719316005706787
ResNet Prediction Class: 963 (Confidence: 1.00)
Status: Likely OOD
--------------------------------------------------
Image: /kaggle/input/testset/generated_image_18.png
Caption: a blue dog bed on the floor
PCS Score: 0.3557380437850952
ResNet Prediction Class: 434 (Confidence: 0.30)
Status: Likely In-Distribution
--------------------------------------------------
```

Figure 5: PCS Score that flags Likeliness

The figure 7 shows PCS (Perceptual Concept Shift) scores and ResNet classification results for two images. The first, captioned "a pizza with pepperoni and pepperoni on it," has a low PCS score (0.307) but a high-confidence prediction (Class 963, Confidence: 1.00), indicating a semantic mismatch and likely Out-of-Distribution (OOD) status. The second image with caption "a blue dog bed on the floor" also has a higher PCS score of 0.356 and lower confidence

(Class 434, Confidence: 0.30), showing closer match between image and caption and is Likely In-Distribution. This again highlights PCS's capability to evaluate semantic grounding over model confidence.

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 misleading out-of-distribution inputs through our proposed Semantic Confusion attack. This method challenges models using visually plausible but semantically contradictory images generated from unnatural captions (e.g., "a zebra rug in a living room") using a diffusion-based generative pipeline. Our working hypothesis was that even well-trained CNNs are susceptible to breakdowns in semantic reasoning when exposed to coherent yet counterintuitive concepts outside their training distribution.

The results supported this hypothesis. Images generated through Semantic Confusion significantly degraded both models' predictive accuracy, with ResNet-18 being particularly brittle. It often displayed high confidence in semantically mismatched predictions, revealing poor alignment between visual input and internal representation. EfficientNet-B0, while also prone to errors, exhibited higher uncertainty on confused samples, as reflected by increased prediction entropy and a greater distributional spread across classes. This indicates that its compound-scaling design may contribute to better semantic calibration under stress, albeit not completely overcoming the challenge.

To systematically evaluate semantic coherence under these attacks, we introduced the Perceptual Concept Shift (PCS) metric. PCS combines Grad-CAM attention alignment, BLIP-generated region captions, and CLIP-based image-caption similarity to quantify how much a model's reasoning deviates from the intended concept. ResNet-18 consistently showed higher PCS scores, indicating greater misalignment between its focus and the true semantic context of the input. EfficientNet-B0, in contrast, exhibited more consistent attention and image-text coherence, though still susceptible to specific hybrid compositions.

Despite these insights, several limitations remain. First, our study focused solely on image classification with CNNs, excluding newer architectures like vision transformers or multimodal systems that integrate richer semantic priors. Second, the Semantic Confusion pipeline depends on curated caption prompts and fixed-generation settings, which may not fully capture the diversity or unpredictability of real-world conceptual shifts. Lastly, while the visual realism of diffusion-generated images is high, certain edge cases may still contain artifacts or implausible details that could bias model behavior or human evaluation.

Future work should address these challenges by integrating adaptive caption generation using reinforcement learning or human feedback to improve the contextual plausibility of prompts. Extending this approach to broader tasks—such as VQA, medical diagnosis, or autonomous perception—could evaluate model reasoning under more domain-specific semantic stressors. Furthermore, incorpo-

rating real-world feedback loops, longitudinal training observations, and multimodal evaluations (e.g., vision-language grounding) may provide deeper insights into how models interpret, adapt to, or fail under semantic dissonance.

0.4 Conclusion

In this work, we introduced Semantic Confusion, a generative adversarial strategy designed to probe the conceptual robustness of vision models by exposing them to visually realistic but semantically misleading inputs. By leveraging language models to produce counterintuitive captions and synthesizing corresponding images with diffusion models, we created a suite of challenging out-of-distribution samples that stress-test a model's semantic reasoning capabilities. We also proposed the Perceptual Concept Shift (PCS) metric, which triangulates model attention (Grad-CAM), semantic region descriptions (BLIP), and image-text alignment (CLIP) to quantify how internal reasoning diverges from intended meaning.

Evaluations on ResNet-18 and EfficientNet-B0 revealed consistent fragility in both models when confronted with conceptually hybrid images. ResNet-18 frequently exhibited overconfidence in incorrect predictions, while EfficientNet-B0, though slightly more resilient, still struggled with compositional novelty. Higher PCS scores indicated poor semantic coherence, especially in ResNet-18, highlighting the importance of attention alignment and multimodal grounding in improving robustness.

These findings underscore a critical limitation in current vision systems—their inability to maintain stable reasoning under subtle yet semantically contradictory shifts. As AI systems increasingly operate in open-world environments, robustness to semantic confusion becomes essential, particularly in safety-critical domains like healthcare, surveillance, and autonomous systems.

Future research should aim to automate and diversify the Semantic Confusion pipeline, apply it across modalities and architectures, and integrate human feedback to enhance realism and relevance. Overall, this work provides a novel and interpretable lens for diagnosing conceptual vulnerabilities in vision models and guiding the development of more semantically grounded AI systems.