Quantifying Neural Network Robustness: A Data-Centric Analysis of Visual and Label Noise

Abstract (Draft)

This research investigates the reliability of Convolutional Neural Networks (specifically a modified ResNet-18) when subjected to data quality degradation. Unlike model-centric approaches that focus on architecture, this study adopts a data-centric perspective to analyze three specific failure modes: (1) visual corruption (noise vs. blur), (2) label inconsistency (annotation errors), and (3) multimodal conflict (image vs. metadata).

Experimental results on the CIFAR-10 dataset demonstrate that while models are highly sensitive to high-frequency visual noise, they exhibit a measurable "uncertainty signal" (Entropy) when facing incorrect labels during early training epochs. Furthermore, the study identifies a "Shortcut Learning" phenomenon in multimodal fusion, where the model prioritizes low-dimensional metadata over complex visual features, leading to critical failures when data sources contradict.

Table of Contents

1. Introduction

1.1 Context: The reliance of Deep Learning on high-quality, curated datasets.

1.2 Problem Statement: Neural networks often fail silently when deployed on "dirty" real-world data (sensor noise, human labeling errors, conflicting metadata).

1.3 Research Objective: To quantify the degradation of a standard ResNet-18 architecture under controlled data quality shifts.

1.4 Scope: Evaluation limited to the CIFAR-10 benchmark using a modified architecture for low-resolution inputs.

2. Methodology and Architecture

2.1 Dataset: Overview of CIFAR-10 and preprocessing steps (Normalization/Un-normalization).

2.2 Model Architecture:

Implementation of a custom ResNet-18.

Specific modification: Replacing the initial 7x7 convolution/max-pool with 3x3 convolution to preserve spatial dimensions for 32x32 images.

2.3 Experimental Framework:

Design of the modular noise-scale injection pipeline.

Hardware setup (MPS/Metal Acceleration on macOS).

3. Experiment I: Visual Robustness Analysis

3.1 Hypothesis: Models degrade differently under "Texture Corruption" (Gaussian Noise) versus "Structural Corruption" (Gaussian Blur).

3.2 Experimental Setup: Injecting graded levels of noise ($\sigma \in [0.0, 0.2]$) and blur ($\sigma \in [0.0, 2.5]$) at inference time.

3.3 Results:

Comparative decay curves of classification accuracy.

Analysis of model sensitivity to high-frequency pixel variance vs. low-frequency shape loss.

4. Experiment II: Label Noise and Memorization Dynamics

4.1 Hypothesis: Neural networks exhibit higher internal uncertainty (Entropy) when learning from incorrect labels before memorization occurs.

4.2 Method: The "Conflicting Labels" Protocol.

Injecting 20% symmetric label noise (e.g., Dog $\to$ Cat).

Metric: Shannon Entropy of the softmax output distribution.

4.3 Results:

The Confusion Signal: Statistical divergence between Clean and Conflicting data entropy (Mean: 0.43 vs. 0.55).

Temporal Dynamics: Identification of the "Early Learning Phase" (Epochs 4-8) where the signal is strongest before the model overfits/memorizes the noise.

5. Experiment III: Multimodal Feature Dominance

5.1 Hypothesis: In multi-source systems, models may develop a bias toward "easier" data modalities (Shortcut Learning).

5.2 Method: The "Feature War."

Construction of a FusionNet (ResNet-18 Image Encoder + MLP Metadata Encoder).

Training on consistent data; Testing on contradictory data (Image $\neq$ Metadata).

5.3 Results:

Quantification of "Modality Collapse."

- Evidence of metadata dominance: Model followed text labels in 81.6% of conflict cases, ignoring visual evidence.

6. Discussion and Limitations

- 6.1 Synthesis: Data quality (correct labels, consistent metadata) is a more critical determinant of reliability than minor architectural tweaks.
- 6.2 The "Lazy" AI: Discussion on why models prioritize shortcuts (metadata) and memorization over robust feature extraction.
- 6.3 Limitations: Experiments restricted to small-scale images (32x32) and synthetic noise types.

7. Conclusion and Future Work

- 7.1 Summary of Contributions: A framework for detecting bad data using Entropy and quantifying multimodal reliance.
- 7.2 Future Directions: Potential for using the Entropy signal to create an "Auto-Cleaning" training loop.

Robustness Analysis of ResNet-18 on CIFAR-10

Focus: Visual Corruption and Label Noise Dynamics

Introduction and Architectural Baseline

The primary objective of this research phase was to establish a robust baseline for evaluating neural network performance under varying data quality conditions. The study utilizes the CIFAR-10 dataset as the primary benchmark.

To ensure relevance to low-resolution input data (32x32 pixels), I implemented a custom modification of the ResNet-18 architecture (`src/models/resnet.py`). Unlike standard implementations designed for ImageNet, this model replaces the initial 7x7 convolution and max-pooling layers with a 3x3 convolution (stride 1) and removes the pooling layer entirely. This preservation of spatial dimensions prevents the premature loss of feature information critical for small images.

Baseline training (`src/experiments/train_baseline.py`) was conducted using Stochastic Gradient Descent (SGD) with momentum and Cosine Annealing learning rate scheduling. By incorporating standard data augmentations (RandomCrop, HorizontalFlip), the baseline model achieves a target validation accuracy of ~85%, providing a stable control group for subsequent stress tests.

Visual Robustness: Texture vs. Structure

A core component of the research involved quantifying the model's degradation patterns when subjected to adversarial visual conditions. I developed a modular noise injection framework (`src/data/image_noise.py`) capable of generating two distinct categories of corruption:

Gaussian Noise: Simulating sensor noise and "micro-fractures" in pixel intensity (Texture Corruption).

Gaussian Blur: Simulating focus loss and edge degradation (Structural Corruption).

The experimentation pipeline (`notebooks/experiment_robustness.ipynb`) subjects the trained baseline model to these corruptions at inference time. By incrementally increasing the severity (Standard Deviation for Gaussian; Sigma for Blur) and un-normalizing/re-normalizing data on the fly, I established a "decay curve" for model accuracy. This comparative analysis isolates whether the model relies more heavily on high-frequency textures or low-frequency shapes for classification.

Label Noise and Memorization Dynamics

Moving beyond visual artifacts, the research also began investigating "Label Noise"—the scenario where training data contains incorrect annotations. I implemented a framework

(`notebooks/train_memory_dynamics.ipynb`) to inject symmetric label noise (randomly flipping labels) into the training set.

The goal of this experiment is to observe the "Early Learning" phenomenon. The training loop tracks two competing metrics epoch-by-epoch:

Generalization: Accuracy on the clean test set (learning true patterns).

Memorization: The rate at which the model learns to predict the *incorrect* specific labels assigned to the poisoned data.

Preliminary setups indicate a focus on identifying the "Memorization Spike"—the specific epoch where the model stops learning general features and begins "cramming" the noise, effectively overfitting to the corrupted data.

The project has successfully transitioned from infrastructure setup to active experimentation. The codebase is now modularized into clear `src/data` (loaders, noise tools) and `src/models` (architectures) directories, decoupling the experimental logic from the core definitions. The current infrastructure allows for rapid iteration on both visual robustness tests and label-noise learning dynamics.

Automated Error Detection via Entropy Analysis (experiment_entropy.ipynb)

Following the robustness tests, the research pivoted to **Data Quality Assurance**. The hypothesis was that a neural network exhibits measurable "hesitation" when presented with training data that contradicts its learned feature representations (e.g., an image of a dog labeled as a cat). To validate this, I implemented an **Entropy (Shannon Entropy)** metric to quantify the model's internal uncertainty during inference.

Experiment A: The "Lie Detector" (10-Class CIFAR-10

I injected conflicting labels into 20% of the training data across all 10 classes and monitored the entropy distribution. The results demonstrated a statistically significant divergence between clean and corrupted data:

Clean Data (Consistent): Exhibited a "Low Entropy Spike" (Mean: **0.43**), indicating high confidence. **Conflicting Data (The Lie):** Exhibited a "High Entropy Smear" (Mean: **0.55**), with a distribution tail extending towards higher uncertainty. This result proves that the model struggles to minimize loss on mislabeled examples, creating a mathematical "signal" that differentiates valid data from errors without human intervention.

Experiment B: Binary Isolation (Cat vs. Dog)

To refine this signal, I restricted the domain to a binary classification task (Cats vs. Dogs). Even in this simplified environment, the "Confusion Signal" persisted. While the model achieved high confidence on the majority of clean data (Green histogram concentrated at 0.0), the conflicting data (Red histogram) displayed a distinct "heavy tail" and scattered probability mass in higher entropy regions. This confirms that **Entropy is a label-agnostic quality metric**—it detects contradictions regardless of the task complexity.

Temporal Dynamics: The "Early Learning" Phenomenon

To understand *when* the model distinguishes between signal and noise, I tracked the entropy of clean versus conflicting labels across 15 training epochs. The results revealed a critical **"Memorization Window."**

Divergence Phase (Epochs 4-8): As the model learned generalized features, the entropy of clean data dropped significantly faster (reaching **0.78** at Epoch 6) than that of conflicting data (remaining at **0.99**)(these came from second experiment). This lag indicates that the model resists learning patterns that contradict its feature extractors. **Memorization Phase (Epochs 12+):** In the late stages of training, the entropy of conflicting data collapsed to near-zero (**0.069** at Epoch 13). This confirms the deep learning theory that neural networks first learn simple patterns (Generalization) before using their excess capacity to brute-force memorize outliers (Overfitting).

Effective data cleaning cannot be a post-training process. The "Confusion Signal" is strongest during the mid-training phase. A robust "Data-Centric" pipeline must therefore intervene during these early epochs to identify and remove conflicting labels before the model permanently encodes them as truth.

Feature Dominance in Multimodal Fusion (experiment_entropy.ipynb)

To evaluate the model's decision-making hierarchy, I trained a fusion network on both raw pixel data (ResNet-18 feature extractor) and categorical metadata (MLP encoder). The model achieved near-perfect training accuracy, but a "Conflict Test" revealed a critical failure mode.

When presented with contradictory inputs (e.g., Image="Dog", Metadata="Cat"), the model exhibited a massive bias toward the textual modality:

Metadata Reliance: 81.6% (408/500 samples). **Visual Reliance:** 11.2% (56/500 samples). **Stochastic Confusion:** 7.2% (36/500 samples). This demonstrates **Shortcut Learning**. The network minimized its loss function by latching onto the low-dimensional, high-signal metadata while effectively discarding the high-dimensional, noisy visual data. This proves that without

specific regularization (like Modality Dropout), multi-sensor systems will naturally degrade into single-sensor systems, creating a hidden single point of failure.