Approach

This project tackles the task of moderating unsafe image content using deep learning. We fine-tuned a **ResNet-50** model for binary classification — labeling images as **"Safe"** or **"Unsafe"**.

Due to hardware and time limitations, we trained on a reduced dataset of just **200 images** (100 safe and 100 unsafe). To make the most of limited data, we used **transfer learning** by freezing the feature extraction layers of the pretrained ResNet and training only the final classification layers.

We also applied **data augmentation techniques** like horizontal flipping, rotation, and color jittering to increase data variability and improve generalization.

While our current dataset is balanced, the system is also designed to handle **imbalanced datasets** by enabling the use of **weighted loss functions**. This allows for assigning higher importance to underrepresented classes (like “safe” in real-world moderation scenarios), improving the model’s fairness and reliability in skewed datasets.

The final prediction script processes input images and visually annotates unsafe ones with a **red bounding box and confidence label**, while safe images are labeled without alteration.

Challenges

1. **Limited Dataset Size**
   Training on just 200 images was a significant challenge. We mitigated this using **extensive data augmentation** and transfer learning, focusing on extracting maximum value from limited samples.
2. **Avoiding Overfitting**
   With such a small dataset, it was essential to avoid overfitting. We froze the base of the ResNet-50 model to retain generalized visual features and only trained the final layers.
3. **Support for Imbalanced Datasets**
   Though our current dataset is balanced, in real-world use, datasets are often skewed (e.g., 24k unsafe vs. 1k safe). The training pipeline supports **assigning higher class weights** to the minority class, helping the model learn fairly even with disproportionate examples.
4. **User-Centric Output Design**
   We aimed for intuitive feedback: unsafe images get clearly flagged with a red box and label, while safe images are displayed unmarked. This ensures the system is both effective and non-intrusive.

Why It’s Awesome

- **Efficient + Lightweight**
  The project demonstrates that even with limited data and compute, a reliable image moderation tool can be built using smart training strategies.

- **Flexible for Real-World Datasets**
  The model is designed to scale — it supports both balanced and imbalanced datasets through weighted loss tuning.

- **Visual + Explainable Results**
  Instead of black-box predictions, users get annotated images with clear, visible labels and confidence scores.

- **Fully Offline & Privacy-Safe**
  The model runs entirely locally, requiring no internet or third-party API — making it secure for sensitive environments.

- **Expandable**
  The current safe/unsafe binary classifier can be extended to detect specific categories like violence, gore, or explicit content using the same framework.