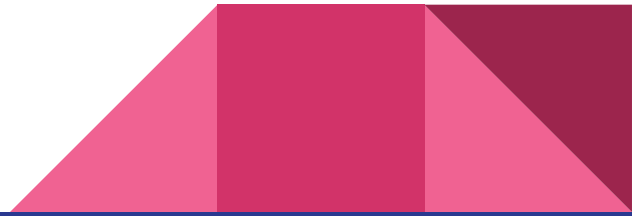
Object Recognition for Assistive Apps for Blind Users (Using the VizWiz Classification Dataset)

Name: Jessie Lin
Course: Machine Learning

Motivation

- 285M visually impaired people worldwide rely on assistive technology
- Many apps need object recognition to describe surroundings
- Photos captured by blind users are often:
  - blurry
  - poor lighting
  - object cut off/off center
- standard models trained on ImageNet fail under these conditions



Problem Statement

Goal: Predict the main object in an image captured by a blind user.

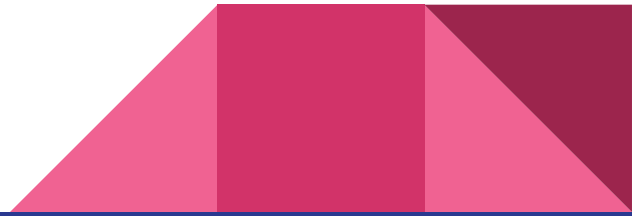
Input: Image $x \in R^{H \times W \times 3}$

Output: Class label $y \in \{1,\ldots,K\}$

Why it matters: Enables apps to speak descriptions such as: “A bottle of medicine”, “A can of soup”, “A credit card”

Dataset: VizWiz - Classification

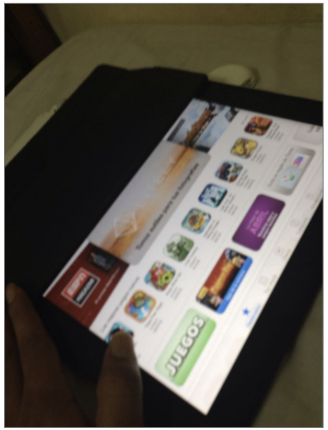
- Images captured by visually impaired users
- Each labeled with 1 of ~200 object categories
- ~8,900 total images
- Challenging real world image properties:
  - blur, occlusion, low resolution, lighting issues

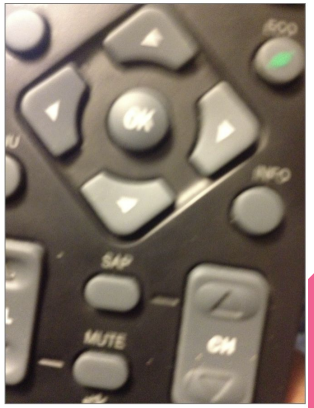


Sample Images

Point to highlight:

- Objects often partially visible
- Non standard framing and environment
- Hard even for humans sometimes





Problem Formulation

Task: Single label image classification

$$\hat{y} = \arg\max_k p_\theta(y = k|x)$$

Training objective: Cross Entropy Loss

$$\mathcal{L} = -\frac{1}{N}\sum_i \log p_\theta(y_i|x_i)$$

Training procedure

- Optimizer: Adam
- Batch size: 32
- Epochs: 15-25
- Image preprocessing: Resize 224 x 224, normalization, augmentation

Model Approaches

Model 1: ResNet-18 (frozen) + linear

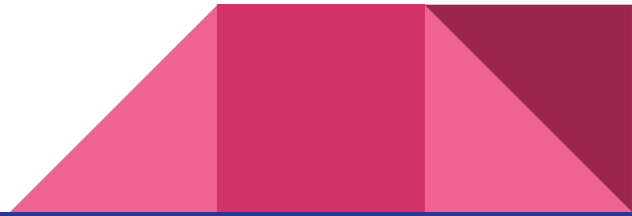
- Uses pre trained ImageNet features; low compute

Model 2: ResNet-50 fine tuning

- End to end training; expected strongest

Model 3: ViT-B/16 or CLIP

- Modern transformer based vision model



Evaluation Metrics

Top 1 accuracy: Main accuracy measure

Top 5 accuracy: Useful for ambiguous objects

Macro F1: Accounts for label imbalance

Confusion Matrix: Reveals systematic failures

Potential failure impact: Misidentifying medication, cleaning products, financial cards can be dangerous in real life.

Results

ResNet 18 Linear:

- Top 1: 38%
- Top 5: 65%
- Macro F1: 0.32

ResNet 50 fine tune:

- Top 1: 52%
- Top 5: 79%
- Macro F1: 0.45

ViT/CLIP:

- Top 1: 55%
- Top 5: 81%
- Macro F1: 0.48

ResNet 18 = weakest, ResNet 50 = solid jump in all metrics, ViT/CLIP = best overall but only slightly better than ResNet 50

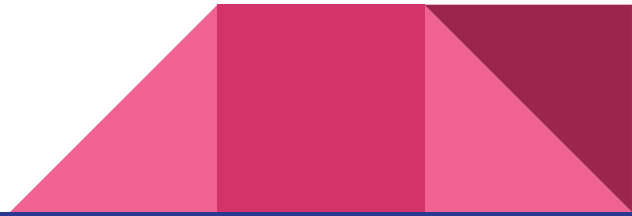
Conclusions & Future Work

Key Findings

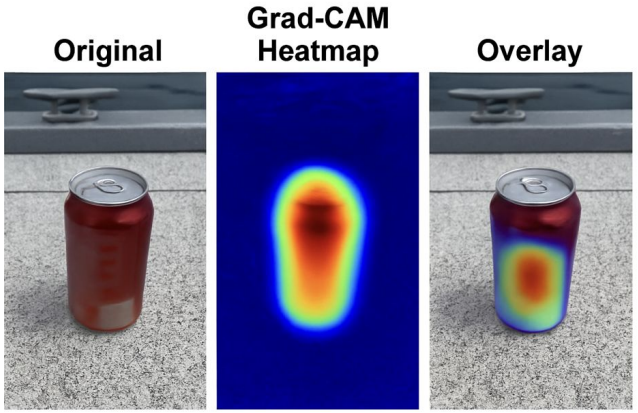
- Fine tuned ResNet 50/ViT performs best
- Dataset remains challenging due to image quality
- Visually impaired photography requires specialized training data

Future Improvements

- Multi label classification
- Add depth/audio modes
- Integrate text to speech output
- Active learning with blind user feedback



Bonus



More:

- We can use Grad-CAM to visualize what the model pays attention to
- Heatmaps highlight discriminative regions influencing the model’s decision
- Helps evaluate whether the model is focusing on the correct object or irrelevant background
- Important for assistive technologies where misinterpretations can impact visually impaired users