SIGNVISION AI

A COMPUTER VISION APPROACH TO VISUAL TRANSLATION AND ITS RISKS

PRESENTED BY:

MIKAYLA NORTON

TABLE OF CONTENTS








BUSINESS OBJECTIVE

"SignVision AI" advances communication for the deaf and hard-of-hearing by developing a computer vision system that uses deep learning to translate sign language alphabet gestures accurately. Addressing the critical challenge of adversarial attacks—deliberate alterations to deceive AI—this research emphasizes the importance of creating robust, secure models. Through experiments and adversarial training, the project enhances the reliability of AI systems, ensuring their practical utility in sensitive applications like sign language recognition.

OBJECTIVES

BASELINE MODEL TRAINING

Initial deep neural network training of sign language alphabet dataset

ADVERSARIAL ATTACKS

Baseline model attack using the fast gradient sign method (FGSM) to generate adversarial samples

ADVERSARIAL TRAINING

Utilization of adversarial samples to retrain the baseline model and increase adversarial robustness

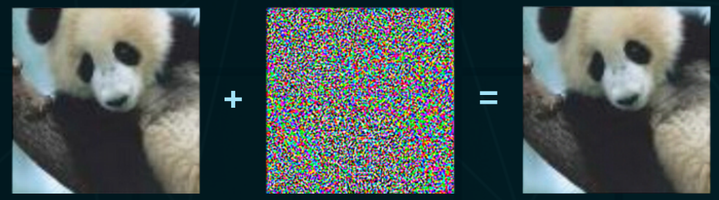
ROBUSTNESS EVALUATION

Exploration of natural and adversarial robustness trade-offs and impacts

ASL ALPHABET

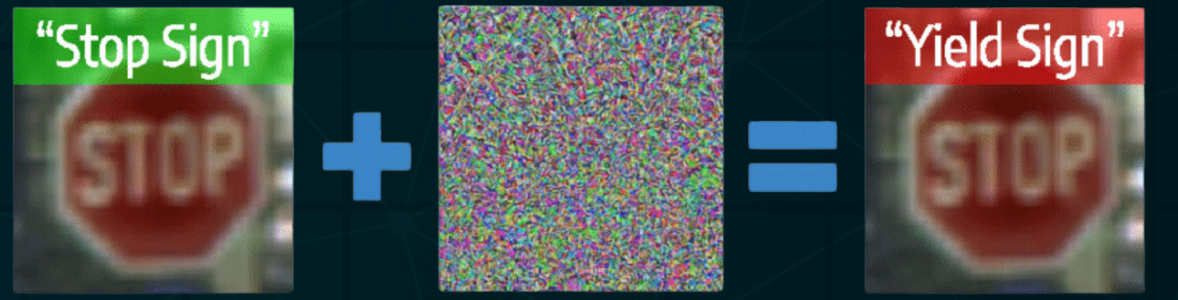
The dataset features a wide variety of hand gestures against a uniform white background, ensuring a consistent visual context for model training. Each gesture corresponds to a letter in the ASL alphabet, and additional gestures for functional commands such as "space" and "delete." The dataset is organized into 29 distinct classes, each represented by 3,000 images, totaling 87,000 labeled images for supervised learning.

ADVERSARIAL ATTACKS

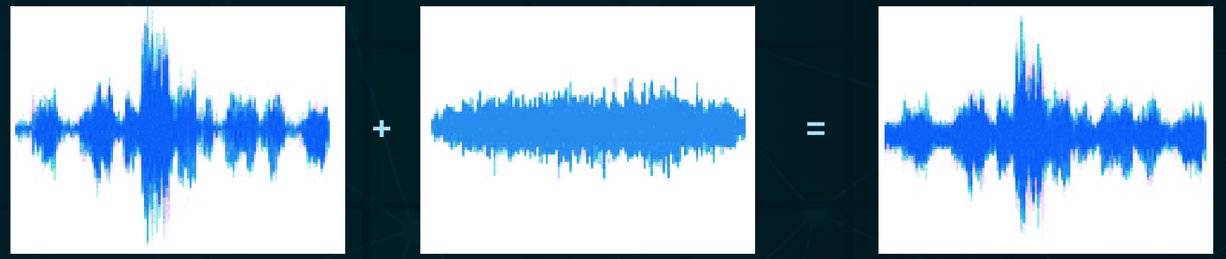


Adversarial attacks involve subtly altering input images in ways that are often imperceptible to humans but cause the model to make incorrect predictions. These attacks exploit vulnerabilities in the model's understanding of input data, leveraging the model's own algorithms against it.

PRIOR WORKS



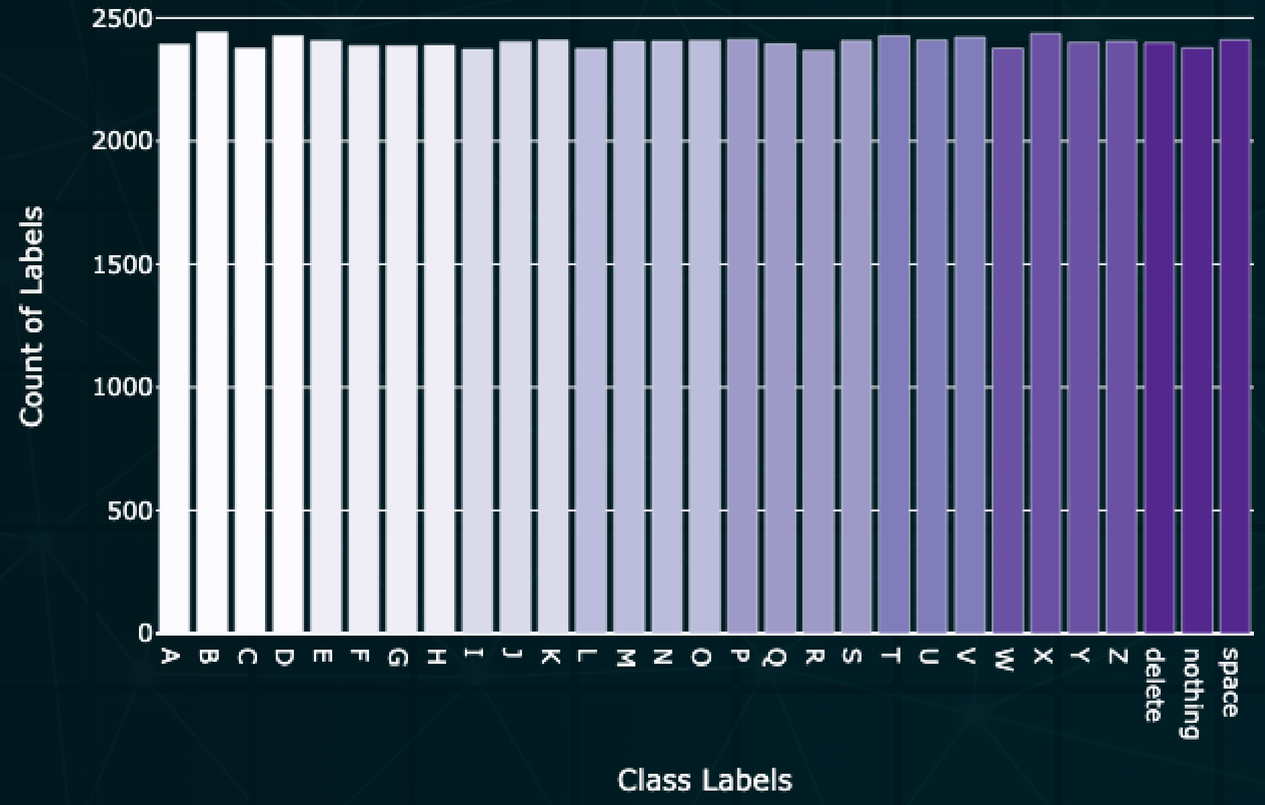
Authentic Input + Adversarial Perturbation = Adversarial Input



“how are you?” + x 0.01 = “open the door”

INITIAL DATA ANALYSIS

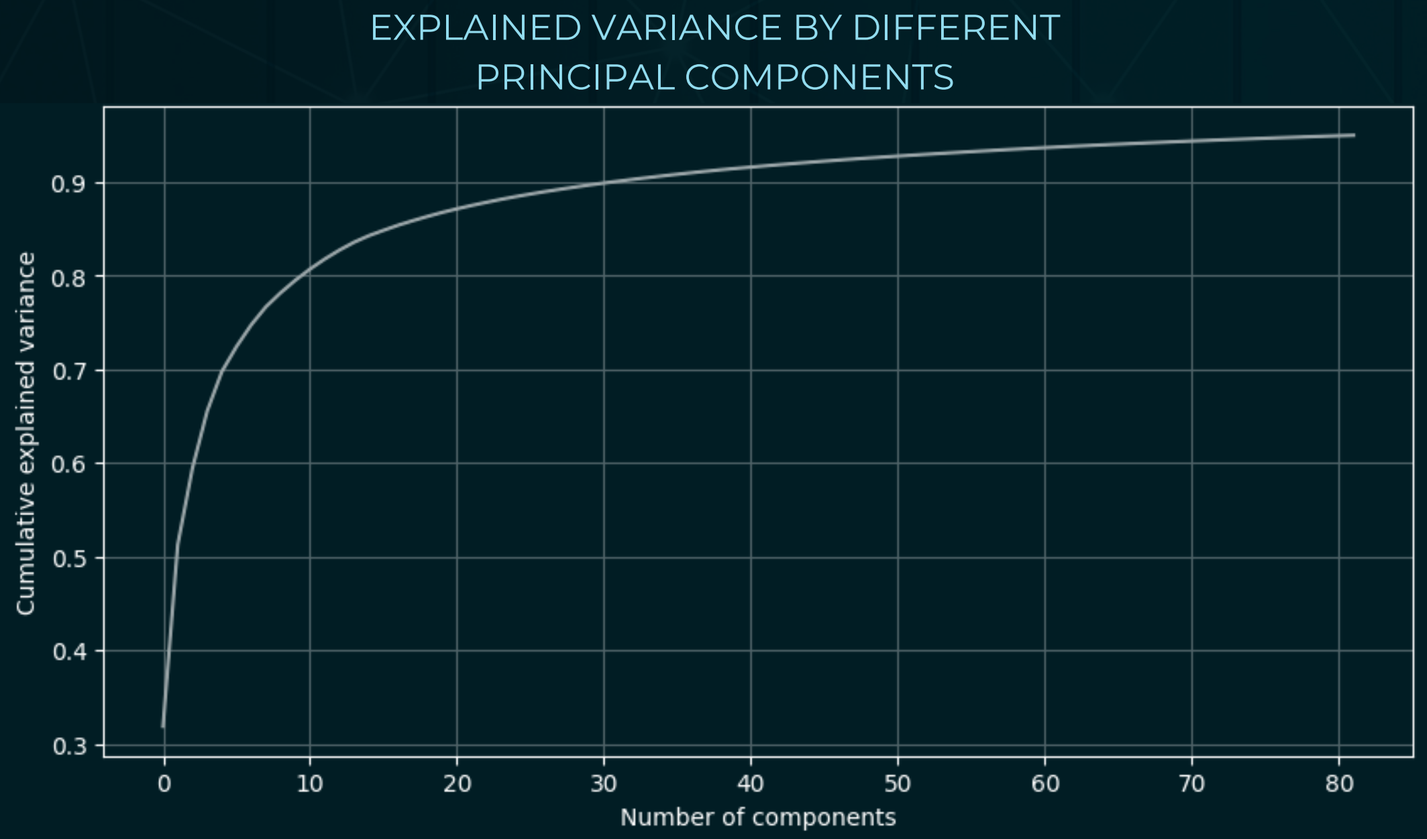
TRAINING DATA CLASSES



29 CLASSES

- 87,000 data points: 3000 imgs/class
- training: ~2400 imgs/class

EXPLORATORY DATA ANALYSIS



PCA+SCALING

- Efficient Dimensionality Reduction: 10-15 PCs
- Optimal Component Selection:
  - balance between complexity and information retention around 10-15 PCs

Settled on 95% variance selection throughout experiment

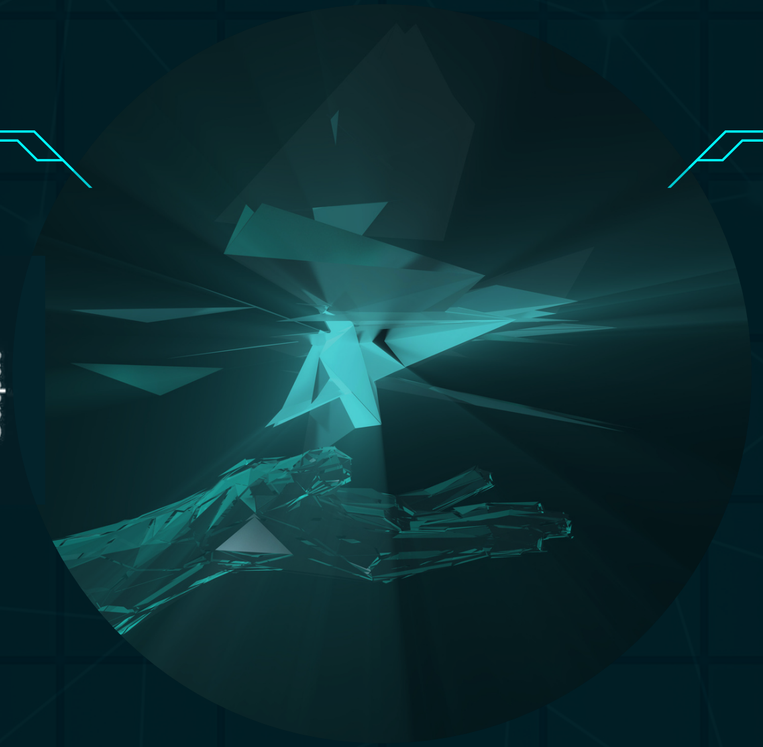
MODEL COMPARISON

The condensed model was significantly more time and memory efficient compared to its full dimension, at the cost of lower accuracy

BASELINE MODEL TRAINING

PRETRAINED MODEL

trained on a large dataset of images and can recognize a wide range of features, output layer modified for unique use-case



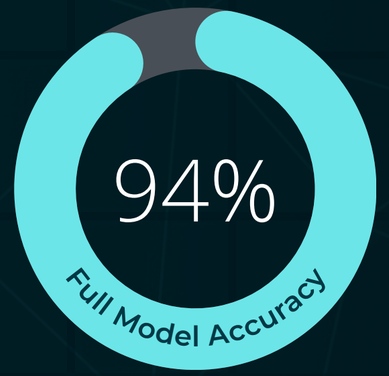
DIMENSIONS

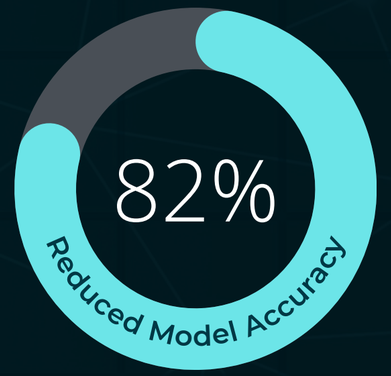
Experimentation with dimensionality reduction shows some tradeoff when utilizing scalers and PCA, compared with near-full dimensional analysis. Images also have low consistency, resulting in significantly lower success with reduced model.

Selection for project: Full Model

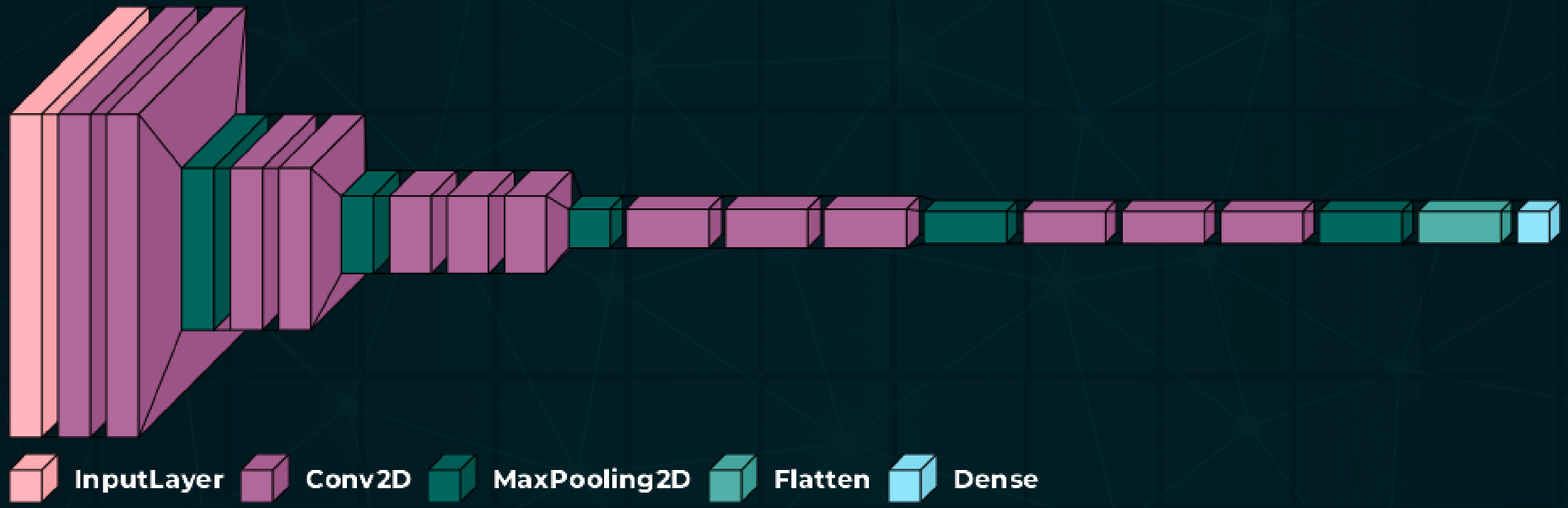
SUCCESS CRITERIA

accuracy and error rates are crucial to consider when evaluating any form of translation model, whether NLP or CV

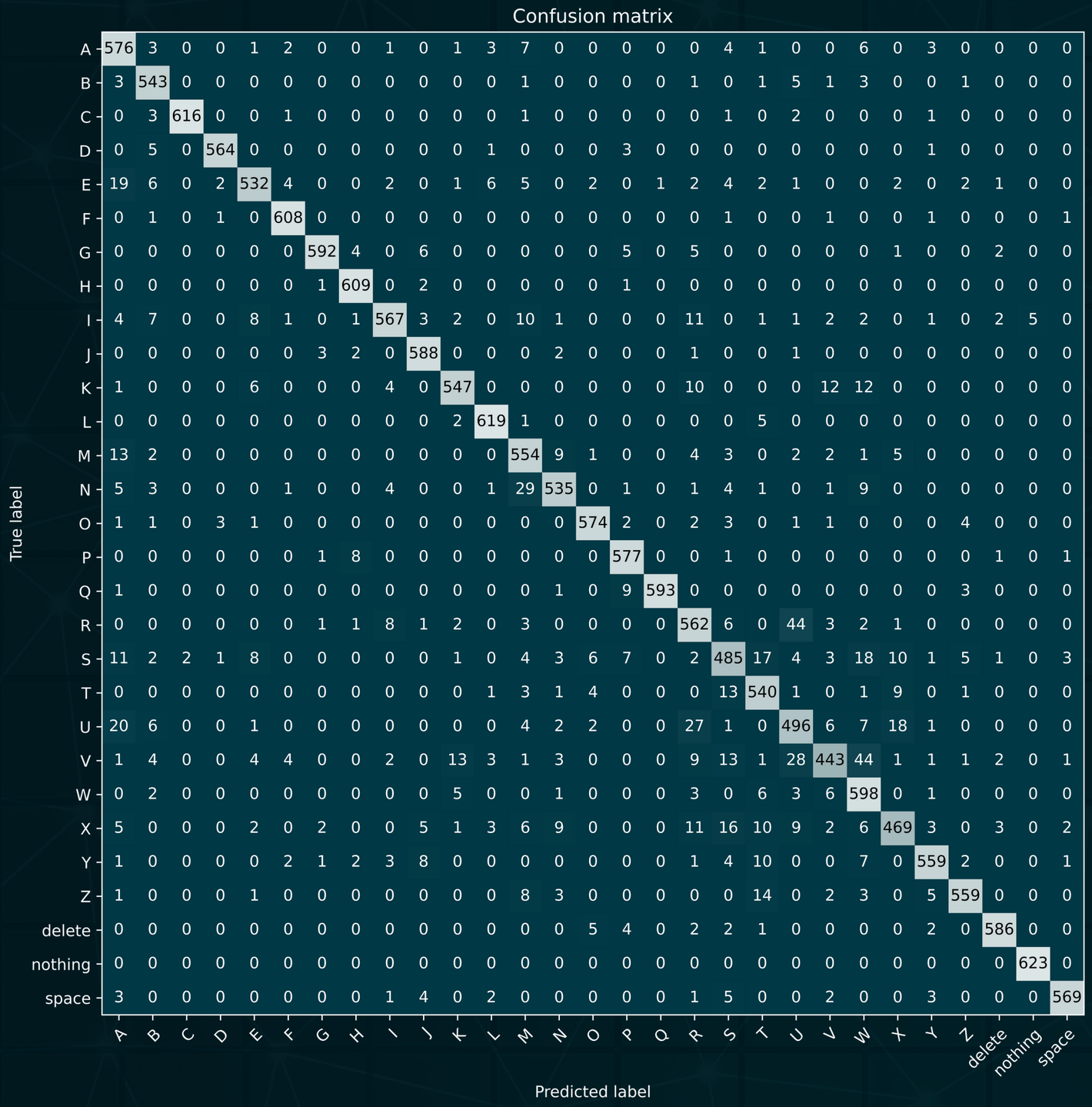




IMPLEMENTATION



Softmax Activation Function	*Categorical Crossentropy* Loss Function	*Adam Optimizer*	Training Epochs n=10	Validation Split 30%



BASELINE RESULTS

Training Accuracy: 94.06%
Validation Accuracy: 93.7%
Testing Accuracy: 93.6%

*some interesting trends between similar letters and confusions

ADVERSARIAL ATTACK TRAINING

WHITE BOX ATTACK

Complete knowledge of the model being targeted

- Attack Method Selection
- Training Setup
- Adversarial Example Generation
- Success Rate Analysis
- Prediction and Confidence Scoring

FGSM

FAST GRADIENT SIGN METHOD

Input Image, x, Label, y, + Target Model, $f(x)$

Loss Function
$J(x,y)$

Gradient Calculation
$\nabla_x J(x,y)$

Perturbation by Sign of Gradient
$\eta = \epsilon \operatorname{sign}(\nabla_x J(x,y))$

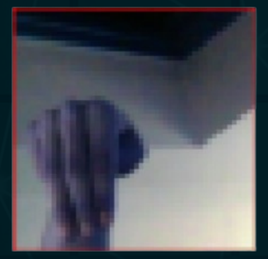
Adversarial Example
$x' = x + \eta$

IMPLEMENTATION

PYTHON'S FOOLBOX

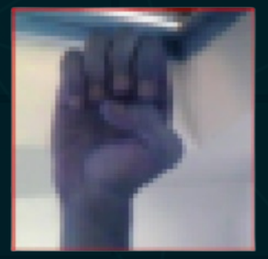
Layers of imperceptible noise are added to each image to generate the adversarial example

Original Images

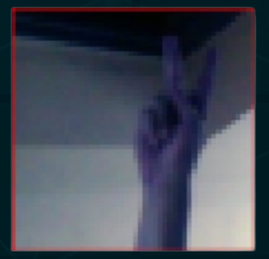


m

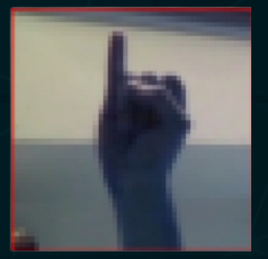
f



e



v



i

Noise
(amplified to maximum color thresholds)

s

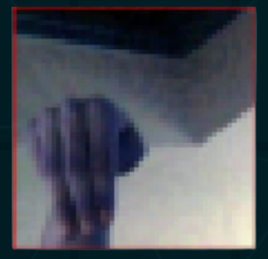
z

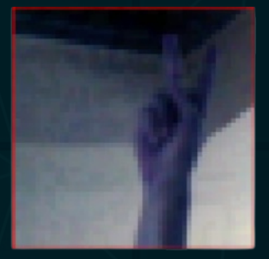
a

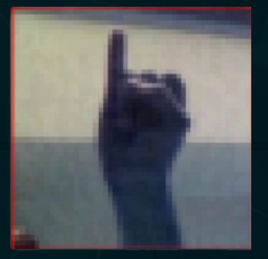
w

k

Adversarial Samples

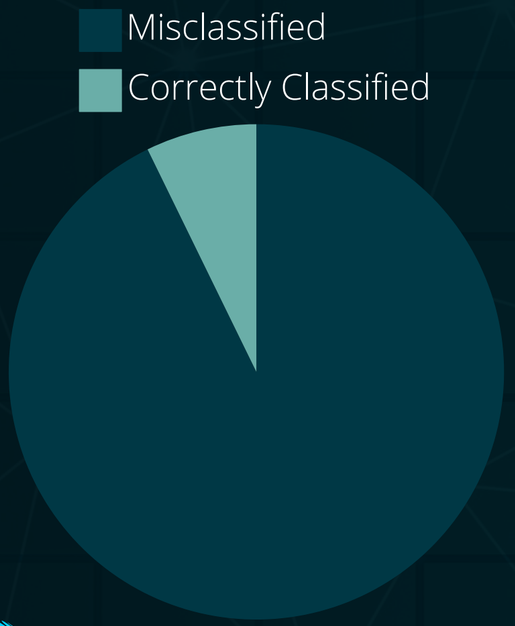






ATTACK RESULTS

AVERAGE ATTACK SUCCESS



Adversarial batch size: 50
Number of batches: 87
Total attacked images: 4,350

Attack Success Range: 90%-100% per batch
Average Attack Success (misclassification rate): 92.76%

(on reduced model: average attack success: 98.81%)

ADVERSARIAL ROBUSTNESS TRAINING

MODEL INITIALIZATION

ADVERSARIAL ATTACK

BATCH/EPOCH PROCESSING

ADVERSARIAL GENERATION

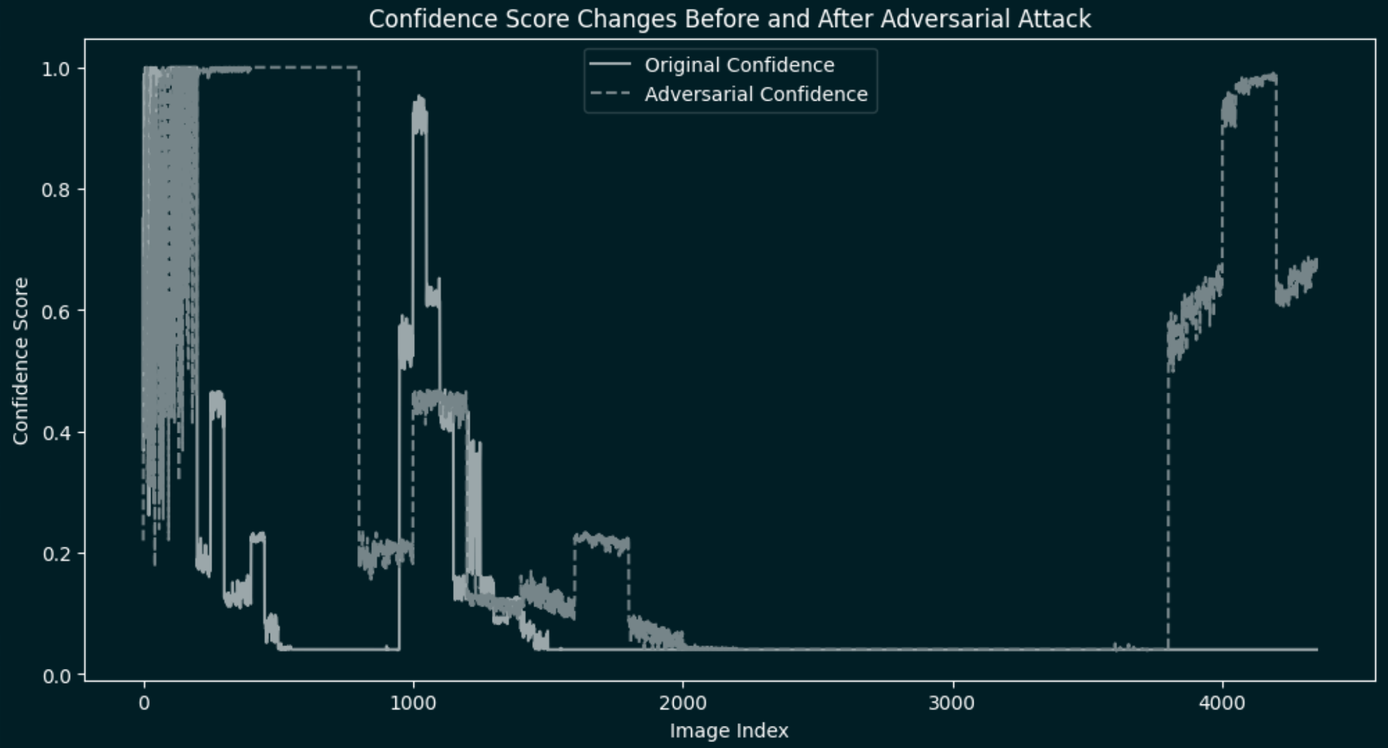
ATTACK EVALUATION

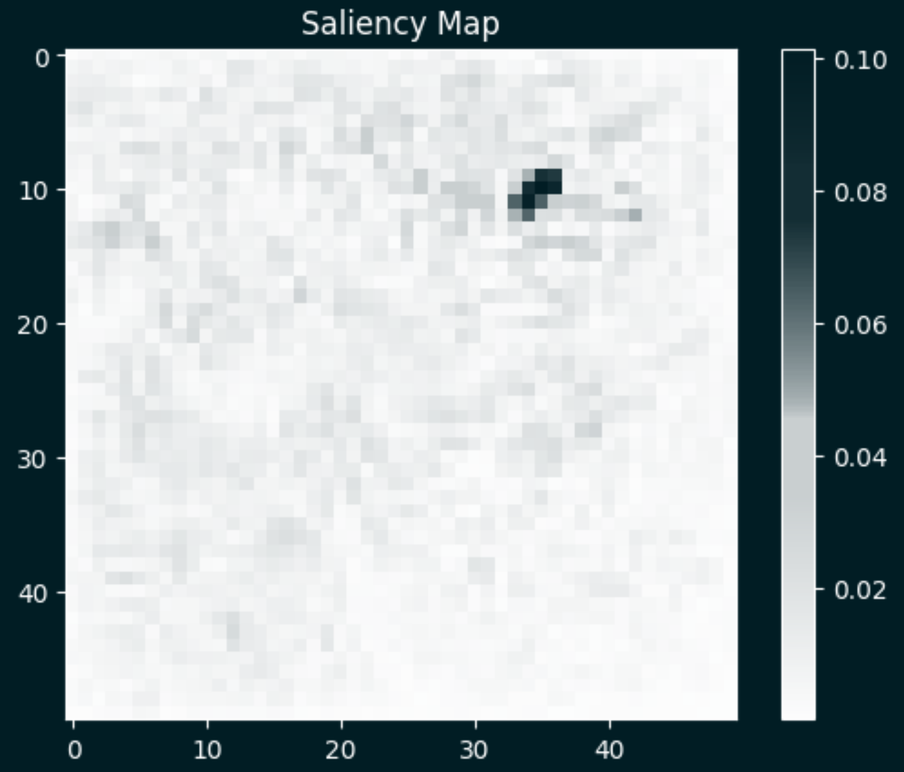
COMBINATION DATASET

NEW MODEL TRAINING

EVALUATE & REPEAT

IMPLEMENTATION

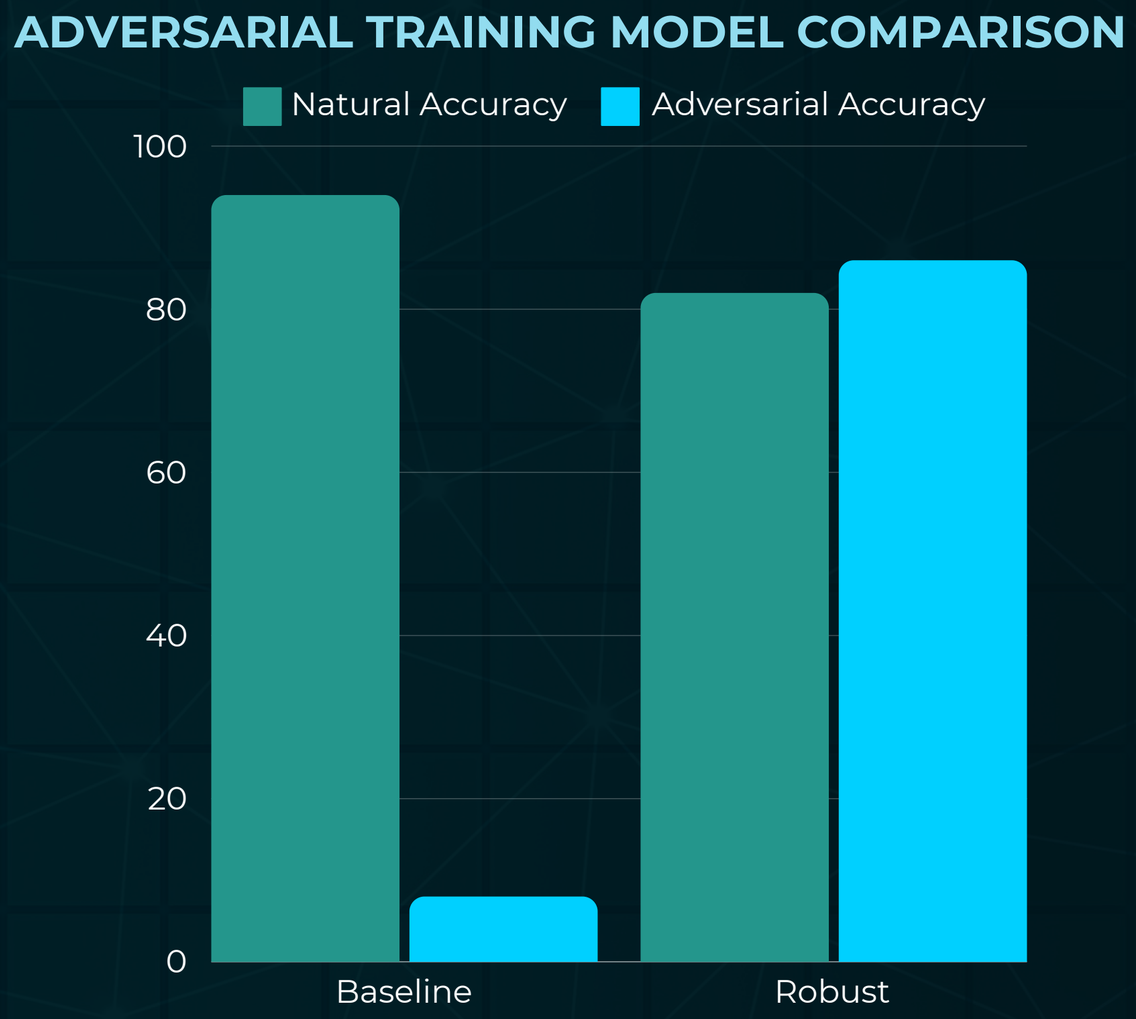




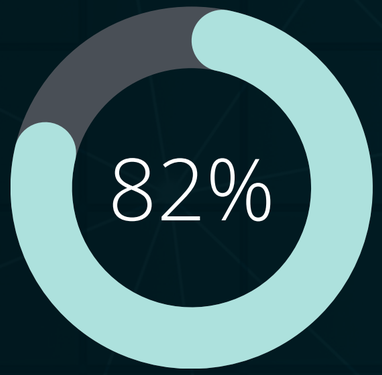
No obvious trends between confidence and success or saliency between images, due to the variant nature of images. However, training was successful.

ROBUSTNESS RESULTS

The adversarially trained model lowers natural accuracy against unperturbed images, however, results in a drastic increase in adversarial accuracy, leading to a more robust model overall.

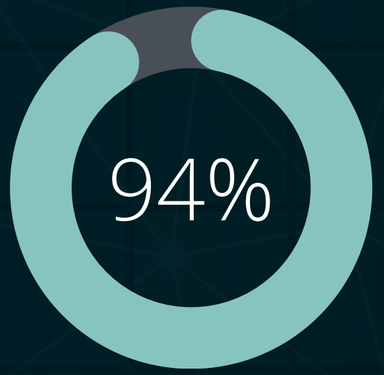


RESULTS SUMMARY



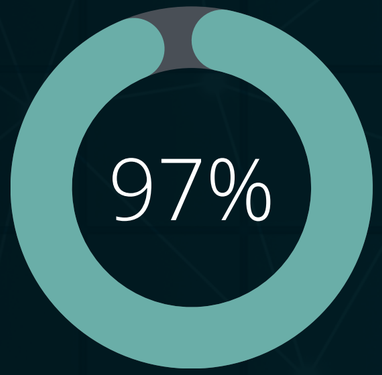
REDUCED MODEL

82% accuracy with computational efficiency but low adversarial robustness



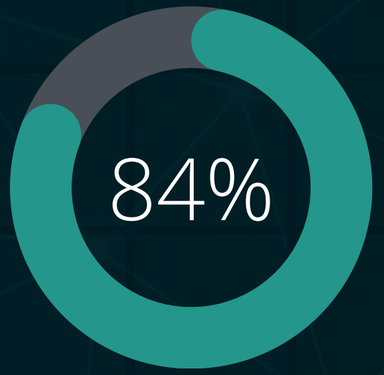
BASELINE MODEL

94% accuracy with higher natural robustness but more computationally expensive



ADVERSARIAL ATTACK

FGSM generates batch samples 90% misclassified or higher by either model



ADVERSARIAL TRAINING

Re-trained full model creates a lower natural accuracy but higher robustness to adversarial attacks

IMPACTS AND CONCLUSIONS

DIMENSIONALITY TRADEOFF

Speeds up computational processes by reducing the complexity of data, but at the cost of losing some information, decreasing the accuracy.

ADVERSARIAL VS NATURAL SECURITY

Improving adversarial robustness leads to decreased natural performance, as enhancing defenses against adversaries can inadvertently reduce sensitivity to legitimate variations in the data.

CALL TO ACTION

To ensure reliable and secure translation for the deaf and hard of hearing, it's crucial to prioritize and enhance adversarial robustness in computer vision models used for any sign language translation.








BUSINESS DELIVERY

Final product delivery included an interactive web app with a main product page to take input sign image, process, and classify the sign for usage.

Web application also features backend access for researchers to monitor reinforcement learning and new data additions moving forward.

THANKS!

any questions?