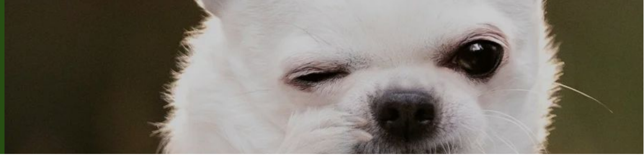
MSDS 634 | SPRING 2024

Happy Doggos: Classifying Dog Expressions

SAMUEL CAMPIONE | IAN DUKE | BELINDA ONG | SONAL SHAD

Hi there! We are team happy doggos!



Sonal Shad

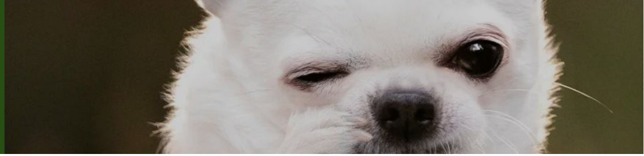
Ian Duke

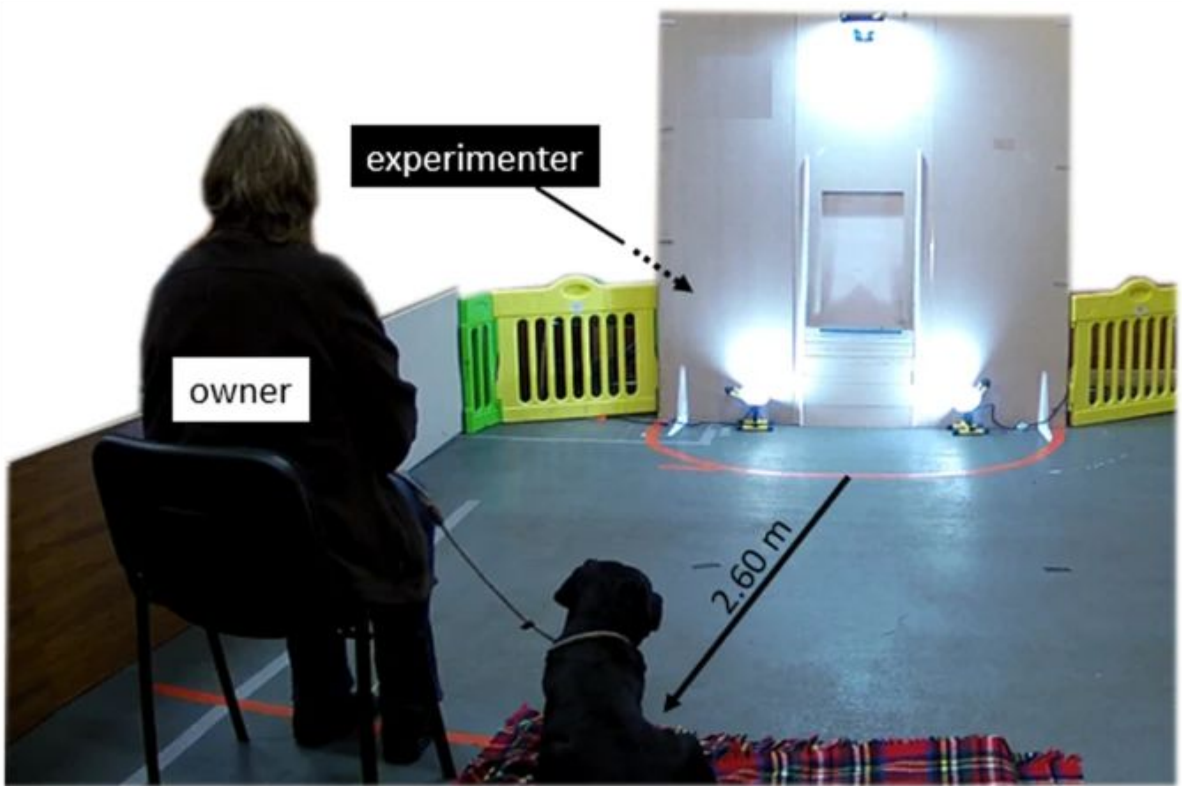
Samuel Campione

Belinda Ong

Problem Statement: We are trying to classify dog emotions based on their facial expressions

Data Collection





Source: Bremhorst A. et al. (2019)

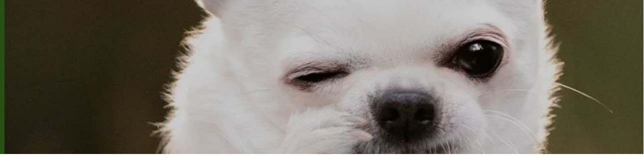
Set-up used is defined by Bremhorst et. al. (2019):

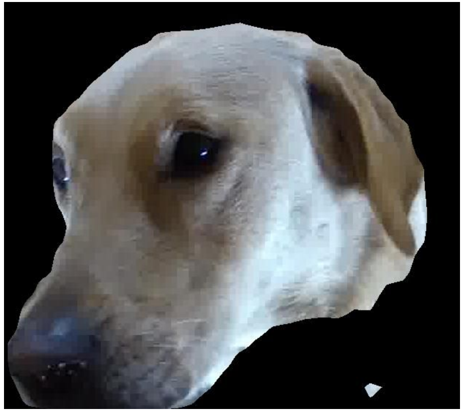
- Experimenter remained hidden behind the wooden wall throughout the experiment
- Food reward is dispensed through the window and they dog's face is filmed

Experimental conditions:

- <u>Positive:</u> One-third of the time, there will be food
- <u>Negative:</u> Two-thirds of the time, there will not be food

Data Preprocessing





Source: Bremhorst A. et al. (2019)

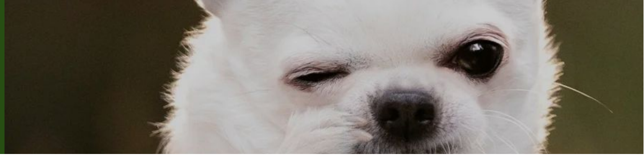
Zamansky's team:

- 18,600 images collected via 248 three-second videos with 25 frames per second
- Identified location of canine face on each 1920 x 1080 frame
- Applied a mask to remove background clutter
- Cropped all images to 500 x 500 pixels
- Manually coded each frame based on experimental conditions

Our team:

- Split data into train, validate, and test dataset
- Created a data loader: batch size 64, shuffled training dataset
- Augmented the data to increase variability and improve model's generalizability through:
  - Random cropping
  - Random rotations
  - Random flips

Model Implementation



Approach

A transfer learning approach was employed using pre-trained models:

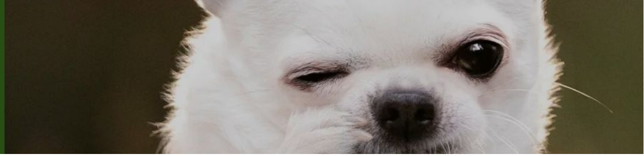
- Convolution Neural Network: ResNet 34 and 50
- Vision Transformer Architecture: DINOv2

These models were chosen as they previously returned promising results in similar tasks. This allows us to benchmark our performance as we experiment with different depths of the ResNet backbone, and newer versions of the vision transformer.

Modification

- The final fully connected layer that outputs the original 1,000 classes in ImageNet was replaced with binary classification of emotional states to output logits for two classes.

Methods



Training

- Using cross-entropy loss, we fine tuned the model over 10 epochs with early stopping.
- Training loss and validation loss and accuracy were calculated in training.

Evaluation

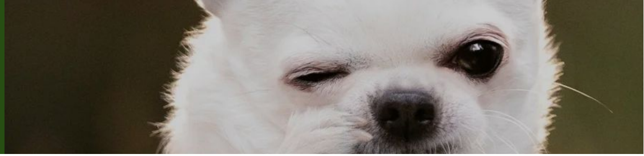
- First, calculated accuracy per frame.
- We took a majority vote across all frames of a given video to determine the predicted label per video.
- Then, we calculated the accuracy, precision and recall per video using the majority vote.

Hyperparameters

We performed a grid search to find optimal hyperparameter values for:

- Learning Rate: [0.0001, 0.001, 0.01]
- Weight Decay: [0.0001, 0.001, 0.01]
- For DINOv2, we additionally tested learning rate 5e-6 as specified in the publication.

Conclusion



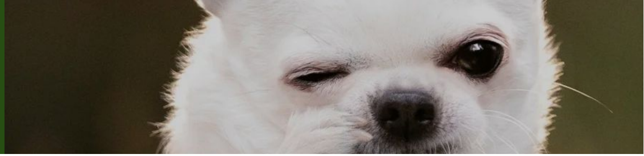
Our findings

- ResNet50 yields validation accuracy of 98%
- ResNet34 validation accuracy 82%
- DINOv2 validation accuracy 89%
- Hyperparameters used in prior research (a learning rate of 10^{-4} for ResNet backbones and 5×10^{-6} for ViT backbone) were suboptimal. Instead, our findings suggest better performance may be achieved employing
  - ResNet34: LR: 0.01, WD: 0.0001
  - ResNet50: LR: 0.01, WD: 0.001
  - DINOv2: LR: 0.01, WD: 0.0001

Original paper

- ResNet50 yielded validation accuracy of 78% (hyperparameter: LR: 0.0001)
- DINO ViT yielded validation accuracy of 85% (hyperparameter: LR: 5×10^{-6})

Conclusion



Test Set Metrics!

- ResNet50 yields test set accuracy of 91%
- ResNet34 test set accuracy 88.9%
- DINOv2 test set accuracy 82.2%

gif-finder.com

HAPPY DANCE