How wrong is wrong? Richer characterization of AI classification error for real world applications

A case study using dog breed classification

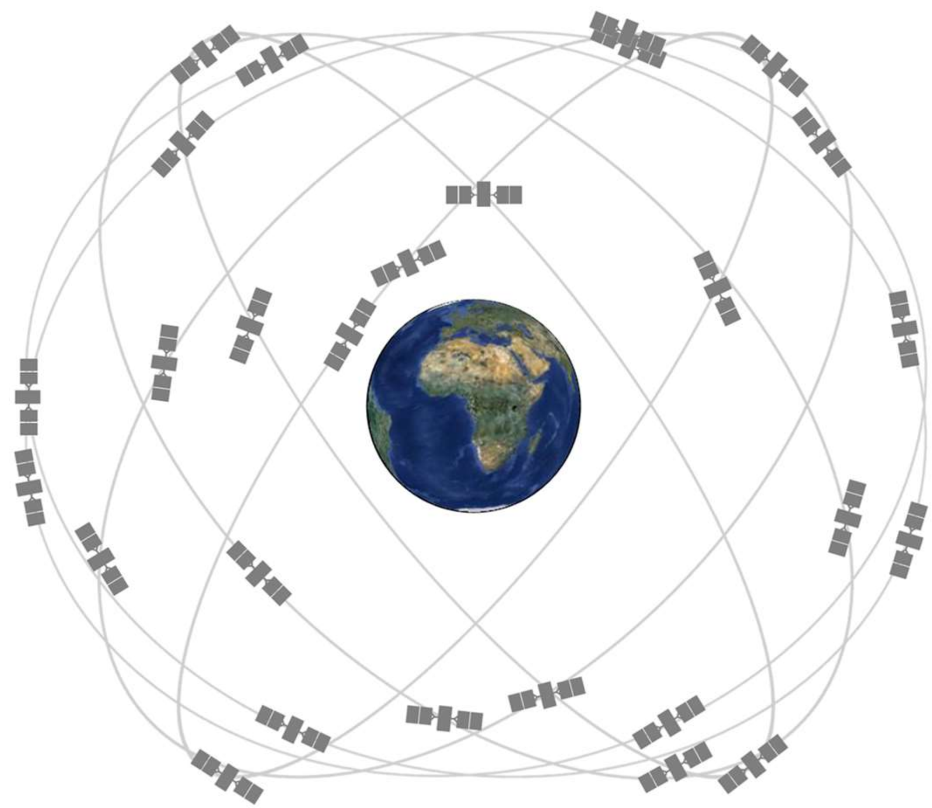
Justine Manning
Joint work with Robert Pless and Zoe Szajnfarber
George Washington University

AI is everywhere

It is being integrated within bigger and more important systems.

There's a disconnect between the way AI engineers *evaluate performance* and the *needs for certification* of complex systems.

What are the failure modes?



Source: https://www.gps.gov/systems/gps/space/

For example, image classification for targeting

Friend or foe?

When the model is inaccurate, what kinds of mistakes is it making?

Source: https://www.vectorstock.com/royalty-free-vector/gun-crosshair-sight-symbols-vector-2666457

Accuracy is not always enough

The computer science view of evaluation of AI tends to focus on accuracy. This table is an example of a core feature of many Machine Learning papers, where the key numbers are all accuracy on a test dataset.

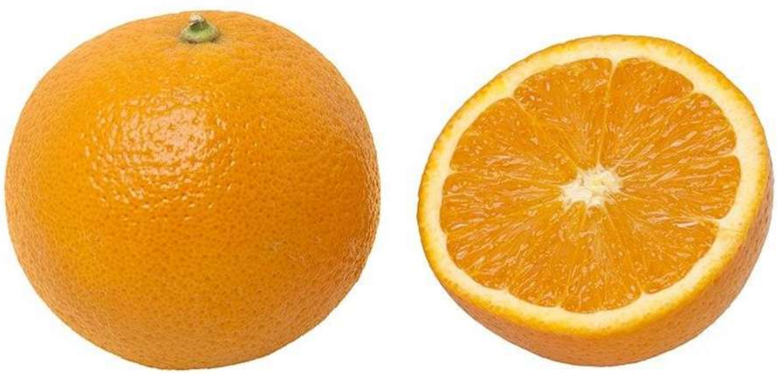
However, to use an AI model, it also important to understand the nature of its errors.

Model	ImageNet	ImageNet-A	ImageNet-R	ImageNet-V2	ImageNet-Sketch	ObjectNet	Average
CLIP [12]	76.2	77.2	88.9	70.1	60.2	72.3	74.3
ALIGN [13]	76.4	75.8	92.2	70.1	64.8	72.2	74.5
FILIP [61]	78.3	-	-	-	-	-	-
Florence [14]	83.7	-	-	-	-	-	-
LiT [32]	84.5	79.4	93.9	78.7	-	81.1	-
BASIC [33]	85.7	85.6	95.7	80.6	76.1	78.9	83.7
CoCa-Base	82.6	76.4	93.2	76.5	71.7	71.6	78.7
CoCa-Large	84.8	85.7	95.6	79.6	75.7	78.6	83.3
CoCa	**86.3**	**90.2**	**96.5**	**80.7**	**77.6**	**82.7**	**85.7**

Table 4: Zero-shot image classification results on ImageNet [9], ImageNet-A [64], ImageNet-R [65], ImageNet-V2 [66], ImageNet-Sketch [67] and ObjectNet [68].

Source: Current best algorithm for ImageNet from Google AI, “CoCa: Contrastive Captioners are Image-Text Foundation Models”, https://arxiv.org/abs/2205.01917

Intuitive Example:



Classifier 1

Classifier 2

Sources: https://commons.wikimedia.org/wiki/File:Orange-Whole-%26-Split.jpg, https://en.wikipedia.org/wiki/Grapefruit#/media/File:Grapefruits_-_whole-halved-segments.jpg, https://commons.wikimedia.org/wiki/File:Banana-Single.jpg

Back to a safety-critical system

It is one kind of error to confuse various type of fighter aircraft and another to confuse a fighter for a passenger airplane when determining friend-or-foe

Source: https://en.wikipedia.org/wiki/Fighter_aircraft , https://en.wikipedia.org/wiki/Boeing_777

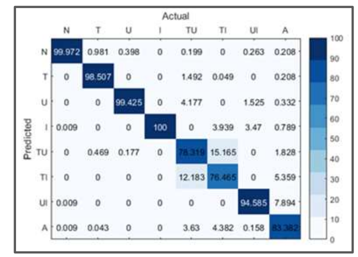
How image classifiers are usually assessed

- Accuracy: of predictions what percent are correct?

Model	ImageNet	ImageNet-A	ImageNet-R	ImageNet-V2	ImageNet-Sketch	ObjectNet	Average
CLIP [12]	76.2	77.2	88.9	70.1	60.2	72.3	74.3
ALIGN [13]	76.4	75.8	92.2	70.1	64.8	72.2	74.5
FILIP [61]	78.3	-	-	-	-	-	-
Florence [14]	83.7	-	-	-	-	-	-
LiT [32]	84.5	79.4	93.9	78.7	-	81.1	-
BASIC [33]	85.7	85.6	95.7	80.6	76.1	78.9	83.7
CoCa-Base	82.6	76.4	93.2	76.5	71.7	71.6	78.7
CoCa-Large	84.8	85.7	95.6	79.6	75.7	78.6	83.3
CoCa	**86.3**	**90.2**	**96.5**	**80.7**	**77.6**	**82.7**	**85.7**

Table 4: Zero-shot image classification results on ImageNet [9], ImageNet-A [64], ImageNet-R [65], ImageNet-V2 [66], ImageNet-Sketch [67] and ObjectNet [68].

- Confusion matrices: what classes are mistaken for each other?



- Visualization: what part of the image contributes to the prediction?

These don’t tell us how bad the error is when the model fails.

Source: https://www.researchgate.net/figure/Confusion-matrix-for-8-class-classification-in-the-SAE-model_fig3_310671661

What can we do with better error information?

1) Choose systems whose error profiles are tolerable in their context

1) Build systems that make fewer bad errors

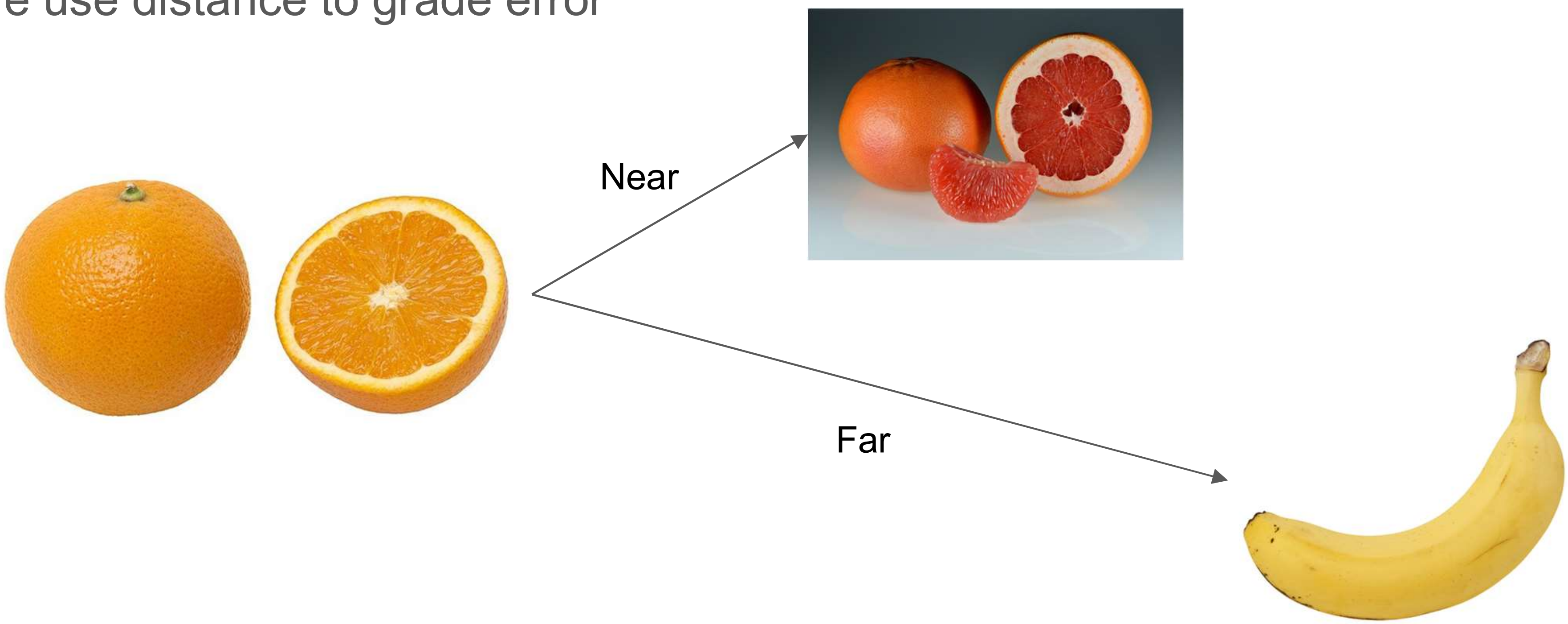
How do we define the magnitude of error?

We have an intuition now for “bad” classifier errors, but how do we actually quantify that?

Our approach makes use of a known and *measurable* hierarchy in the data in order to grade error based on *distance* in the hierarchy.

How do we define the magnitude of error?

We use distance to grade error



Research question:

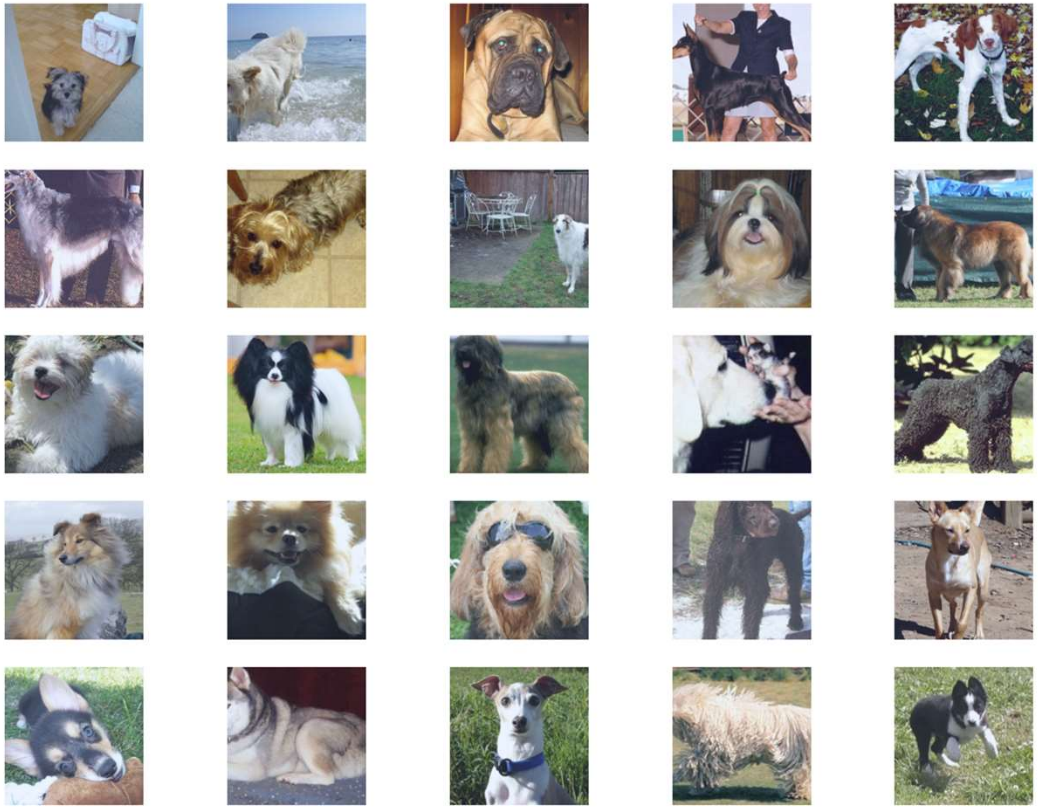
Can we use a hierarchy corresponding to class labels to compare the performance of two image classification models?

We hypothesized that a deeper network would produce “smaller” errors based on the hierarchy than a shallower network.

Case study: classification of dog breeds

To test our hypothesis,

1. We fine-tuned two CNNs to classify images of dogs by breed
2. We built a hierarchy for grading error based on the genetic differences between dog breeds.



The image dataset

We use the Stanford Dogs Dataset, which is intended for fine-grained image classification. (http://vision.stanford.edu/aditya86/ImageNetDogs/)

It contains 20,580 images of dogs from 120 breeds.

The task is hard! Many of the dog breeds look very similar.

n02088364-beagle

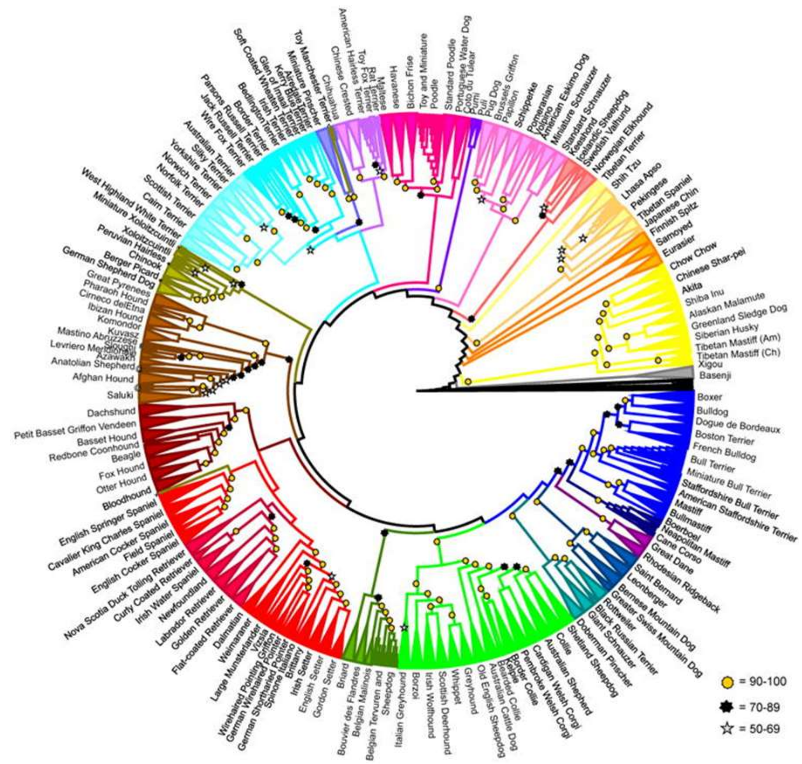
The genetic data

We used data from research about dog genetics.

Parker, Heidi & Dreger, Dayna & Rimbault, Maud & Davis, Brian & Mullen, Alexandra & Carpintero-Ramirez, Gretchen & Ostrander, Elaine. (2017). Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development. Cell Reports. 19. 697-708. 10.1016/j.celrep.2017.03.079.

The genetic data

Parker et al. used dog genetic data to propose a family tree that relates dog breeds to their wild ancestors and each other.



The genetic data

The data provided was distance matrix that showed "the fraction of [genetic] variants that differ between individuals" and the breeds of the individual samples.

We reduced this matrix to the mean distances between breeds

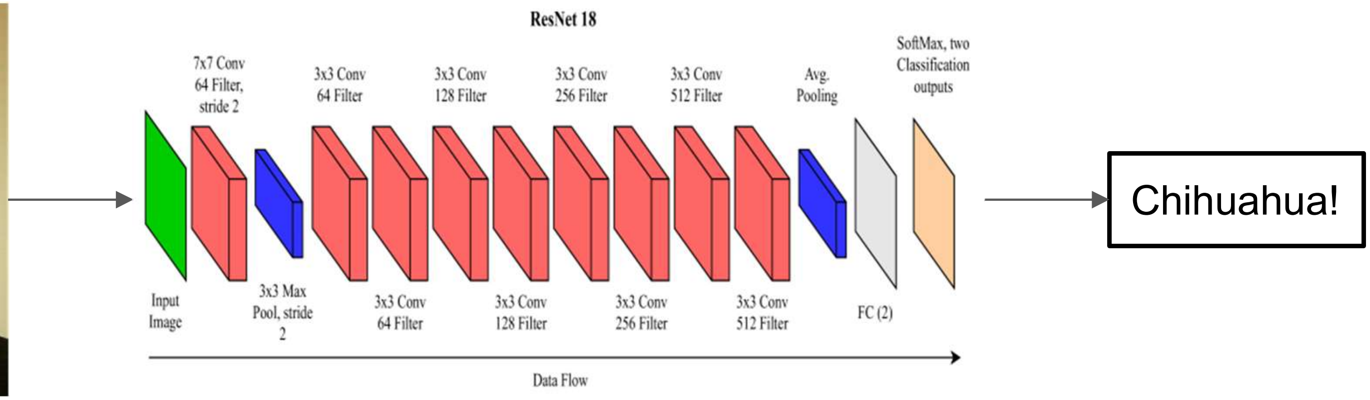
Some values from the matrix:

```
[[0.2661304  0.31064675 0.2935995  ... 0.29895344 0.31096046 0.28942798]
 [0.31064675 0.17679767 0.31368375 ... 0.31757985 0.33137167 0.31165365]
 [0.2935995  0.31368375 0.24351096 ... 0.30026688 0.31363242 0.29514042]
 ...
 [0.29895344 0.31757985 0.30026688 ... 0.25498307 0.30850026 0.2990639 ]
 [0.31096046 0.33137168 0.31363242 ... 0.30850026 0.24997556 0.3104313 ]
 [0.28942798 0.31165365 0.29514042 ... 0.2990639  0.3104313  0.2669295 ]]
```

Min: 0.1513
Max: 0.4126
Mean: 0.3154
Std. Dev.: 0.0170

The models

We fine-tuned pretrained ResNet18 and ResNet 50 models on the same 70% of the data for 20 epochs.



Accuracy on the holdout test set: ResNet18- 76.12%, ResNet50- 89.40%

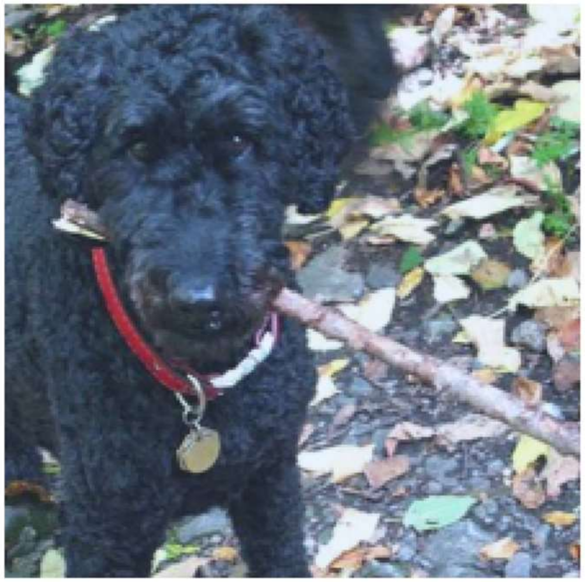
Source: https://deepai.org/publication/tbnet-pulmonary-tuberculosis-diagnosing-system-using-deep-neural-networks

Scoring the error

We recorded the genetic distance between breeds for every misclassification.

Distance: 0.4719

Input: Poodle - Standard



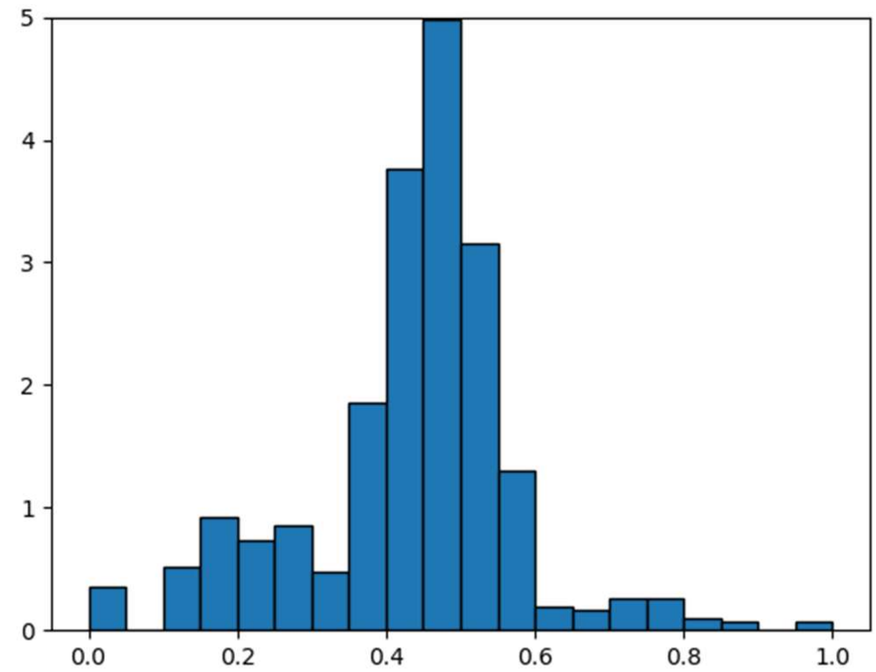
Predicted: Kerry Blue Terrier

Results

ResNet18 density plot

Mean error distance- 0.431

Std. Dev.- 0.145



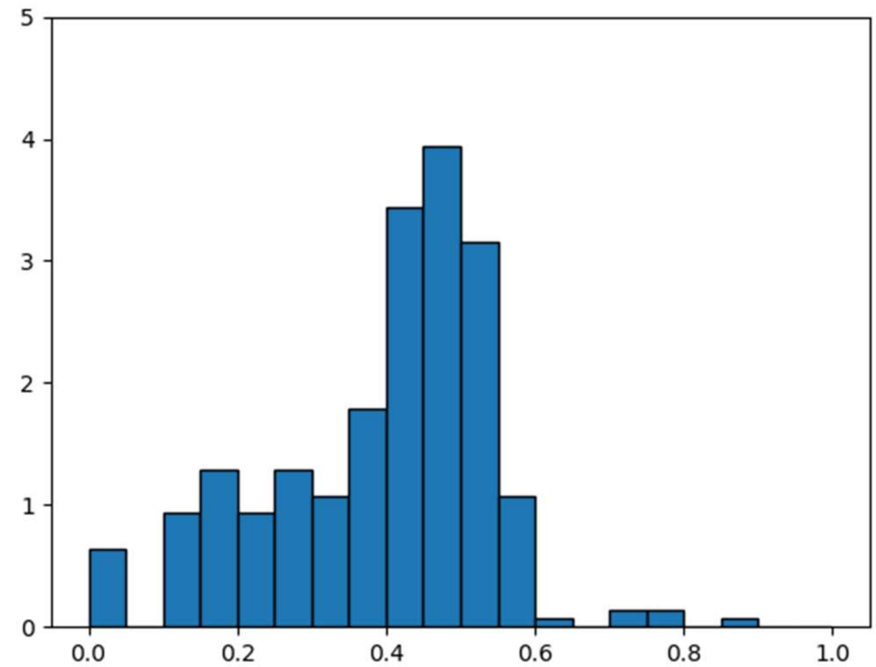
Normalized Distance

Results

ResNet50 density plot

Mean error distance- 0.390

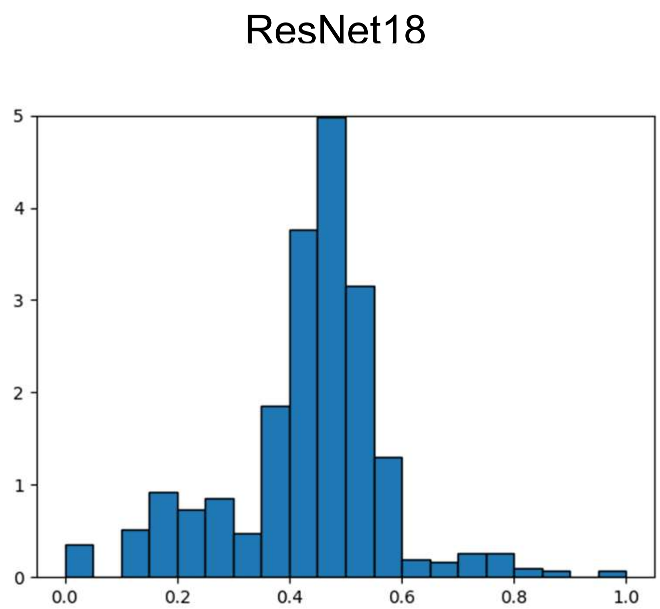
Std. Dev.- 0.154

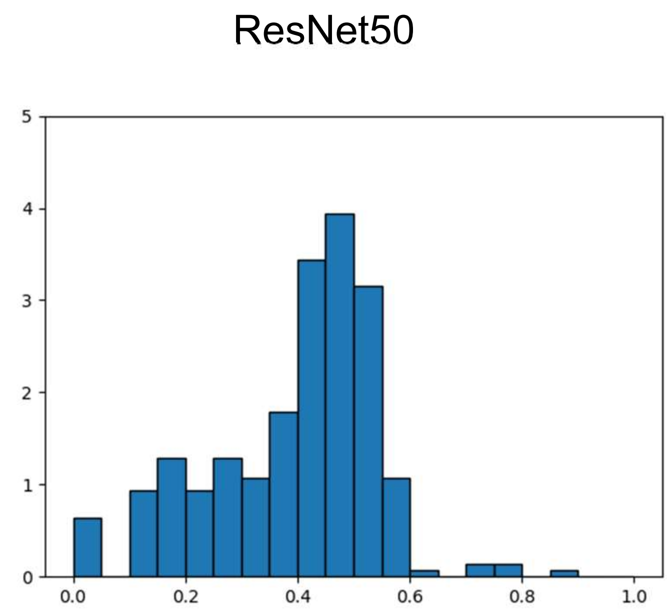


Normalized Distance

Results

In ResNet50, the mean decreased, and the standard deviation increased, so the density of the distribution shifted left and flattened a bit.

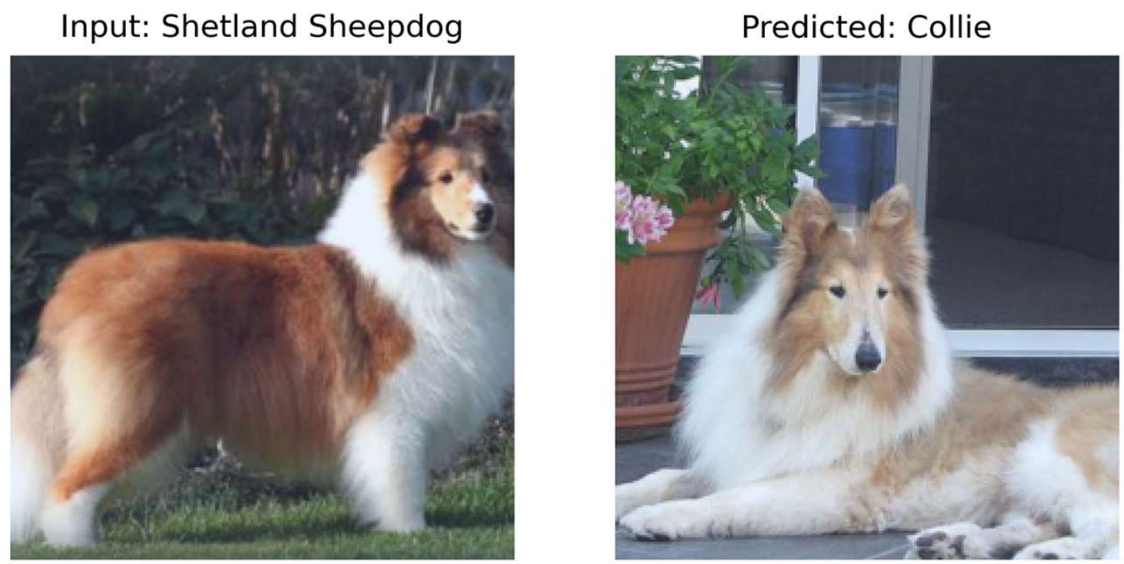




Results

The difference between the distributions was significant (P=0.0002). The lower mean for the ResNet50 network therefore means that the deeper network did lead to smaller errors.

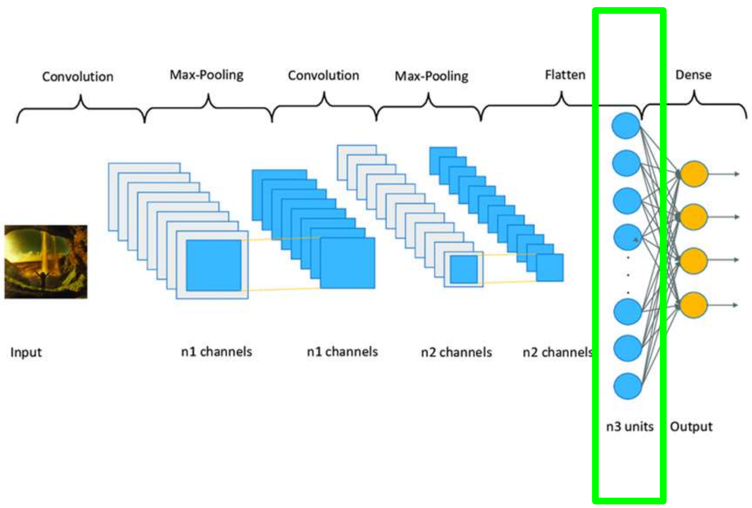
Distance: 0.3258



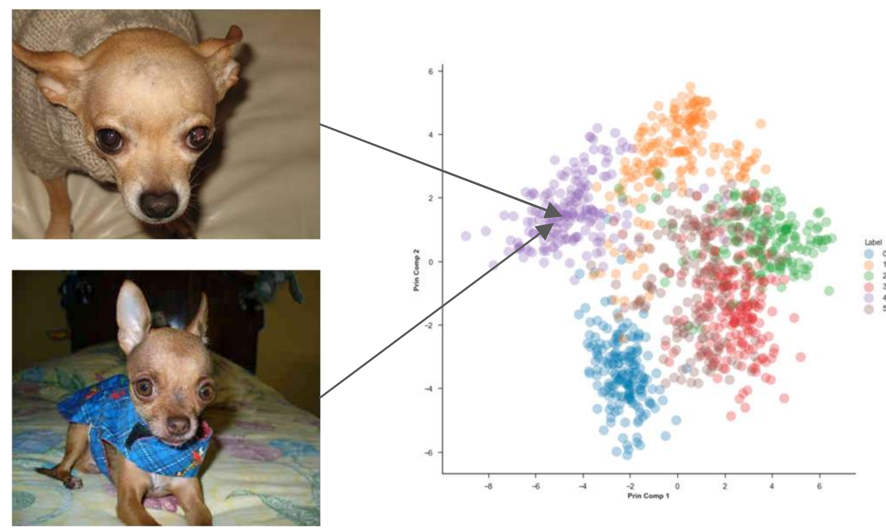
One of the small errors the ResNet50 model made

Combining qualitative assessment of error

We can compare the embedding vectors of different samples.



These vectors show us how the model has placed an image sample in a vector space that allows it to classify the sample.



Source: https://www.researchgate.net/figure/A-vanilla-Convolutional-Neural-Network-CNN-representation_fig2_339447623 , https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1

Model performance - Same inputs to models

ResNet18

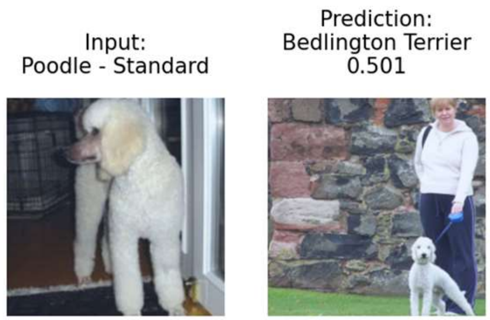
Input:
Poodle - Standard

ResNet50

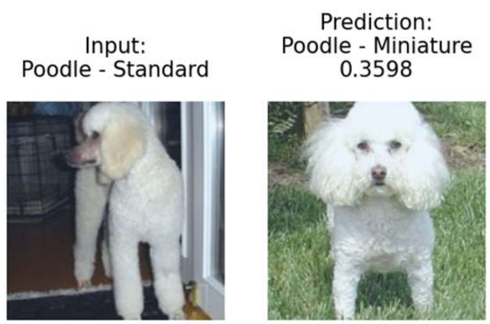
Input:
Poodle - Standard

Model performance - Both models misclassify

ResNet18



ResNet50



Model performance - ResNet50 errors look better

Model performance - Correct breed not in ResNet18 top 5

ResNet18

ResNet50

A richer description of error

By implementing a means of scoring error, we were able to show that ResNet50 is not just more accurate, but also produces smaller errors than the ResNet18 architecture.

Limitations

We used genetic data to compare dog breeds, but genotype is not phenotype. Some dogs that look quite similar are nevertheless distantly related genetically. Confusing them is understandable, but our method scores the mistake severely.

Distance: 0.7215

How to apply this method

The frequency of error distances can be used as metric where a graph structure exists to describe the data.

Where one does not already, an artificial graph could be constructed to suit error-describing needs.

Can the error distance be described as a hierarchy?

This method requires that a hierarchy or distance measure is accessible for comparing the data. For example, classifying objects, what is the distance between one pistol and a smartphone? A way to quantify that distance is required.

?

Source: https://en.wikipedia.org/wiki/Pistol, https://en.wikipedia.org/wiki/Smartphone

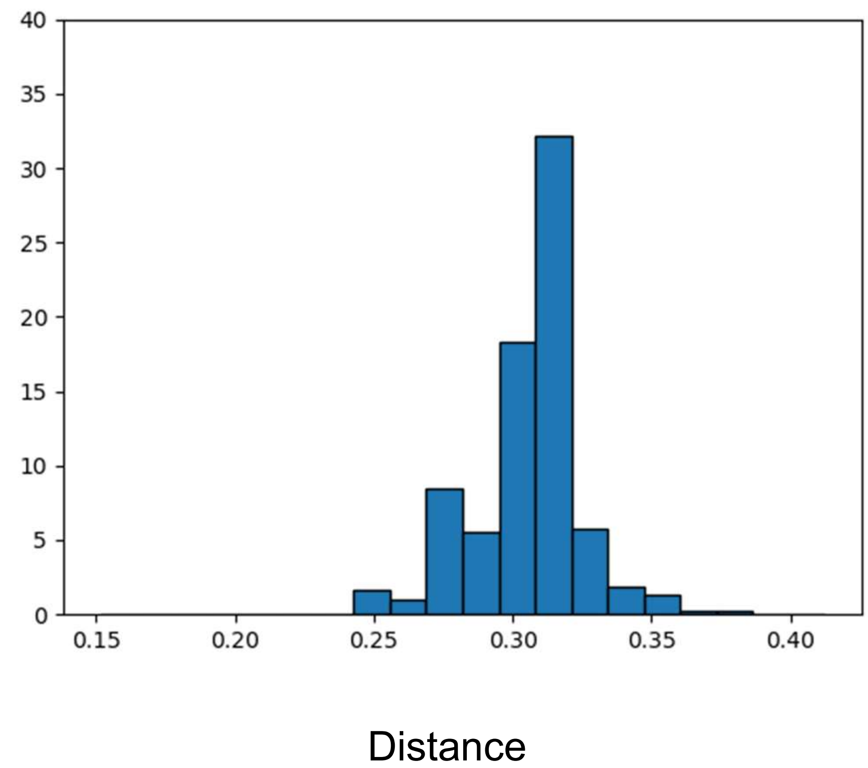
Future research

Using the hierarchy data in the loss function could be a way to push the network into making smaller errors.

Can we incorporate the genetic data during training to push the network to make smaller errors?

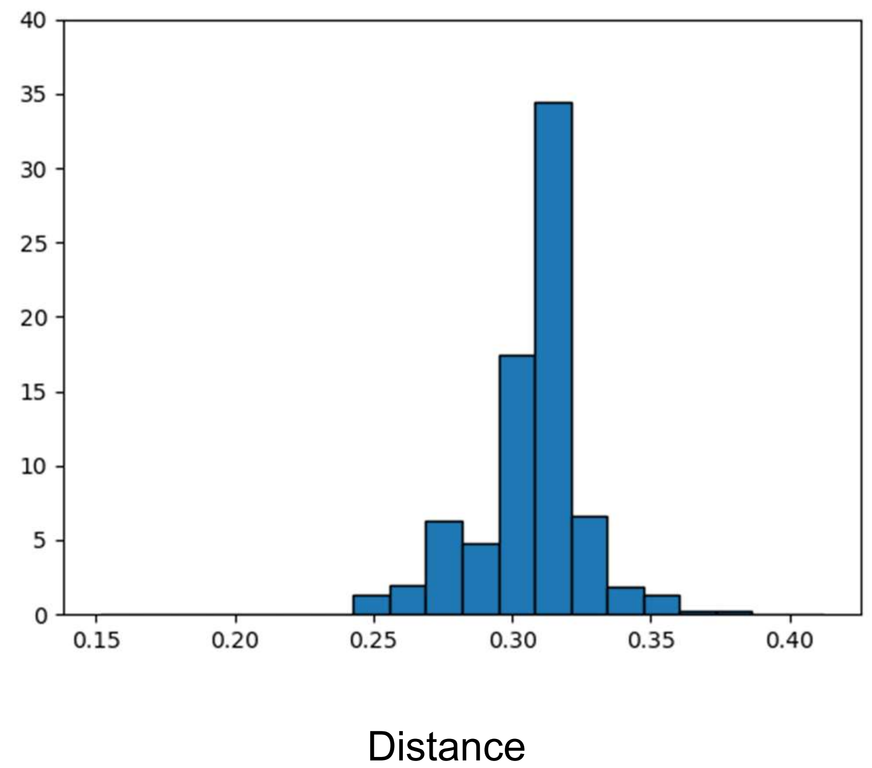
Experiment with distance loss

In ResNet18, using genetic distance for loss pushed the mean down to 0.3057 (P=0.3661) but accuracy went down 3%



Experiment with distance loss

For reference, ResNet18 with categorical loss gave a mean error of 0.3067.



Future research

Further experimentation regarding incorporating the distances into loss is needed, but the initial results are promising.

Thank you for your time!

Please feel free to direct comments and inquiries to

jlsmanning@gwu.edu