CAT BREED CLASSIFICATION

SARA HAMAN

TABLE OF CONTENTS

FINE GRAINED IMAGE CLASSIFICATION

What is Fine-Grained Image Classification (FGIC)?

- Correctly classifying visual objects from ***subordinate categories***, e.g., species of birds, breeds of cats, types of flowers
- "Fine-grained"
  - The challenging aspect
  - ***Small inter-class variations*** and ***large intra-class variations***
- Many state-of-the-art models rely on expensive external annotations
  - Pre-training on large datasets standard for FGIC
  - Dense part (semantic) annotations
  - Recent advances in mixed supervised-unsupervised models (self-attention)
    - **Vision transformers** (NOT CNN)
    - Navigator-Teacher-Scrutinizer Networks
    - Deconstruction and Construction Learning
  - Later two not open-source as of now

FINE GRAINED IMAGE CLASSIFICATION

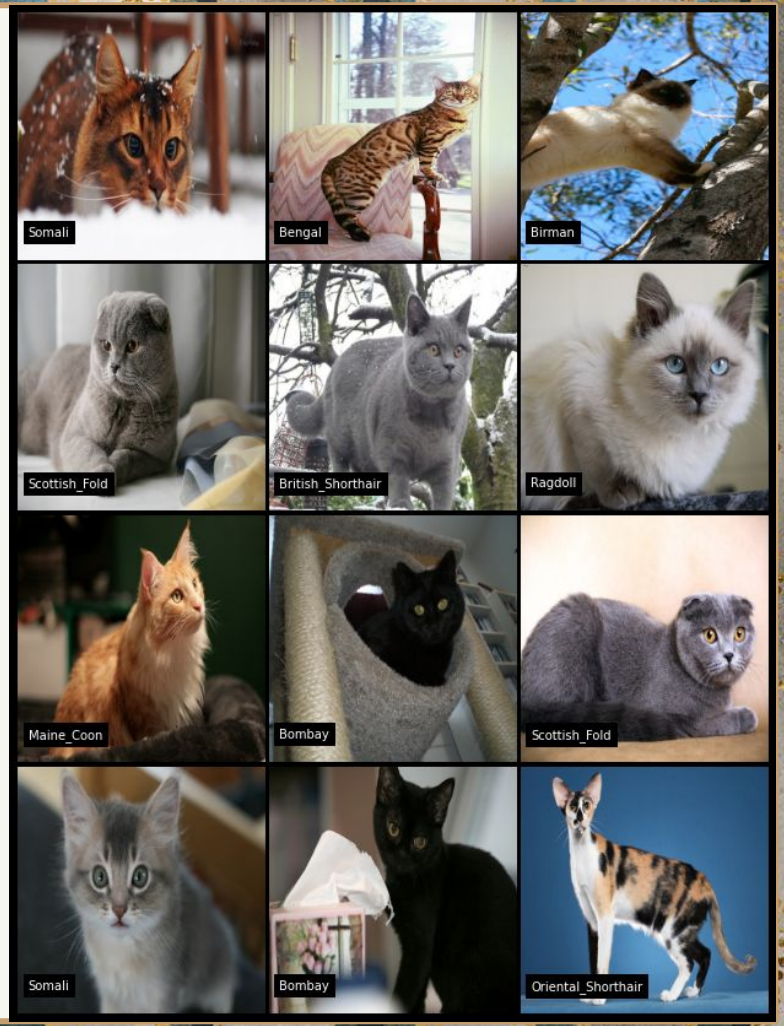
What is Fine-Grained Image Classification (FGIC)?

- Correctly classifying visual objects from ***subordinate categories***, e.g., species of birds, breeds of cats, types of flowers
- "Fine-grained"
  - The challenging aspect
  - ***Small inter-class variations*** and ***large intra-class variations***
- Many state-of-the-art models rely on expensive external annotations
  - Pre-training on large datasets standard for FGIC
  - Dense part (semantic) annotations
  - Recent advances in mixed supervised-unsupervised models (self-attention)
    - **Vision transformers** (NOT CNN)
    - Navigator-Teacher-Scrutinizer Networks
    - Deconstruction and Construction Learning
  - Later two not open-source as of now

THE DATA

1. **Oxford-IIIT Pets Dataset (2012)**
    a. Designed as a ‘challenge’ dataset for ***fine-grained image classification***
    b. Contains images of cats and dogs labeled by their breed
        i. 12 cat breeds, ~200 pictures per breed

2. **Data Expansion**
    a. Collected data for three additional cat breeds
        i. Wanted experience gathering image data
        ii. Selected to better understand how the models learn to identify breeds
        iii. New breeds:
            1. Somali, Oriental Shorthair, Scottish Fold



THE DATA

1. **Oxford-IIIT Pets Dataset (2012)**
   a. Designed as a 'challenge' dataset for ***fine-grained image classification***
   b. Contains images of cats and dogs labeled by their breed
      i. 12 cat breeds, ~200 pictures per breed

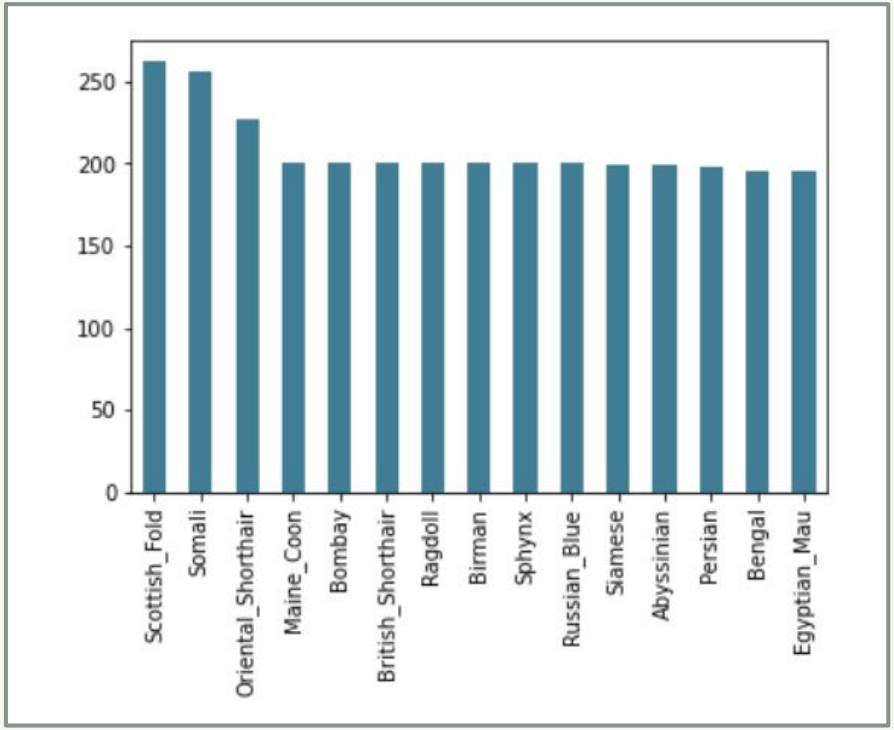
2. **Data Expansion**
   a. Collected data for three additional cat breeds
      i. Wanted experience gathering image data
      ii. Selected to better understand how the models learn to identify breeds
      iii. New breeds:
         1. Somali, Oriental Shorthair, Scottish Fold

DIFFERENT BREEDS!

THE DATA

3. **Final Dataset**
   a. 3,132 images
   b. 15 total cat breeds (classes)
   c. Wide variety of images
      i. Angles, positions, aspect ratios, etc.
   d. Uniformly distributed across classes
   e. Images transformed (random flips and rotations) and standardized (ImageNet mean/std)

   f. HOWEVER,
      i. There are a lot of problems with the data



WHAT IS THE BEST METHOD OF APPROACH?

Differences between cats defined by coat ***color***, body ***shape***, and fur ***texture***?

APPROACH

- **NAIVE CONVOLUTIONAL NEURAL NETWORKS**
  - **“TinyNet”** and **“BabyNet”**
    - Different depths
    - Other minor alterations
  - Trained exclusively on the cat images
  - Not ideal for this problem; hypothesized they would not perform well
- **TRANSFER LEARNING**
  - Networks pre-trained on large datasets (i.e., ImageNet, CIFAR-100)
  - Standard for FGIC tasks
  - **BIG TRANSFER (BiT)**
    - ResNetv2-152 optimized specifically for Transfer learning
  - **VISION TRANSFORMER (ViT)**
    - Not a neural network!
    - Adapted from NLP methods
    - Accounts for positional data

TINYNET

ARCHITECTURE |

- Input size: 3 x 150 x 150
- 2 convolutional layers, 2 dense layers
- 5x5 kernels
- **Optimizer:** SGD
- **Loss Function:** Cross entropy loss

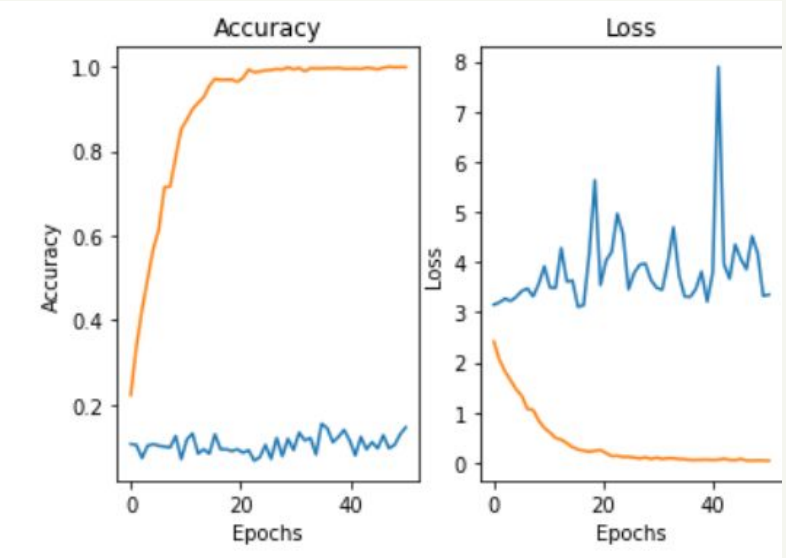
HYPERPARAMETERS |

- … Extensively tuned
- Learning rate: 0.001
- Batch size: 10
- Epochs Trained For: 50

RESULTS |

- Overfit almost right away
- Validation Accuracy: **12.71%**
- Test Accuracy: **14.17%**

TINYNET METRICS



The machine… it does not learn….

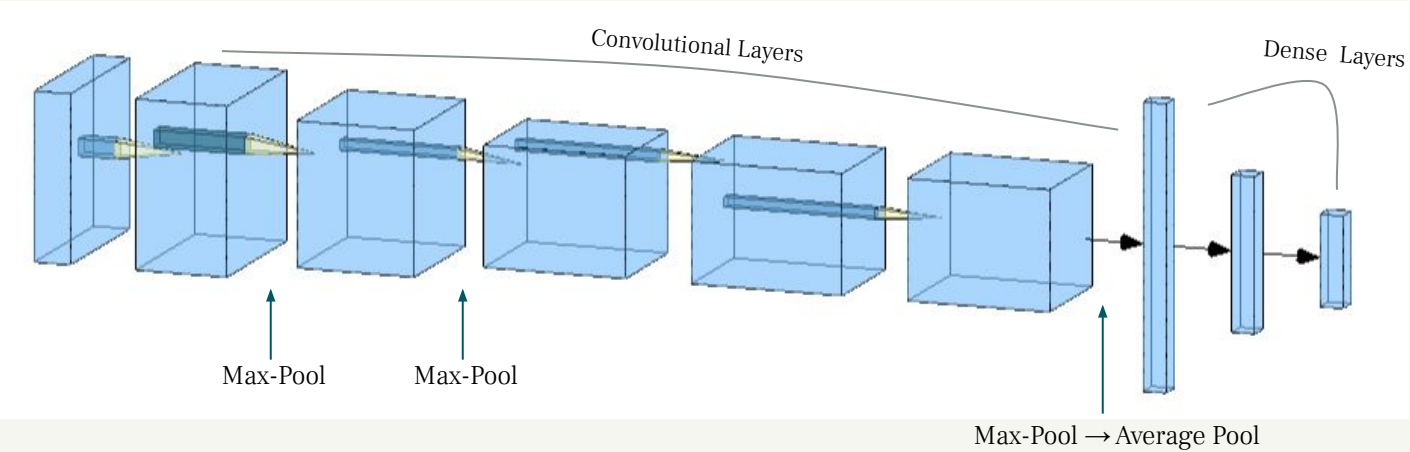
BABYNET

ARCHITECTURE |

- Block structure modeled off of AlexNet
- Batch normalization before every activation
  - Adding dropout improved validation performance
- **LR:** 0.002
- **Optimizer:** SGD
- **Loss Function:** Cross entropy loss

RESULTS |

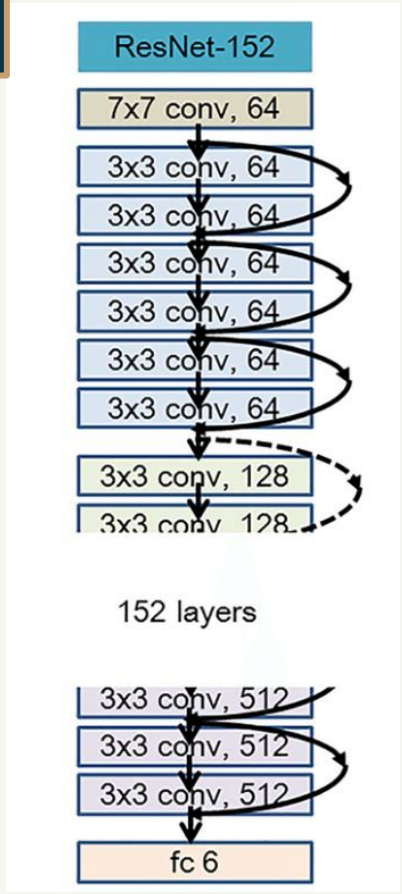
- Does not quickly over-fit to the training data; high levels of regularization
- Even after many epochs, model does not learn anything
- Validation Accuracy: **11.4%**
- Test Accuracy: **12.08%**



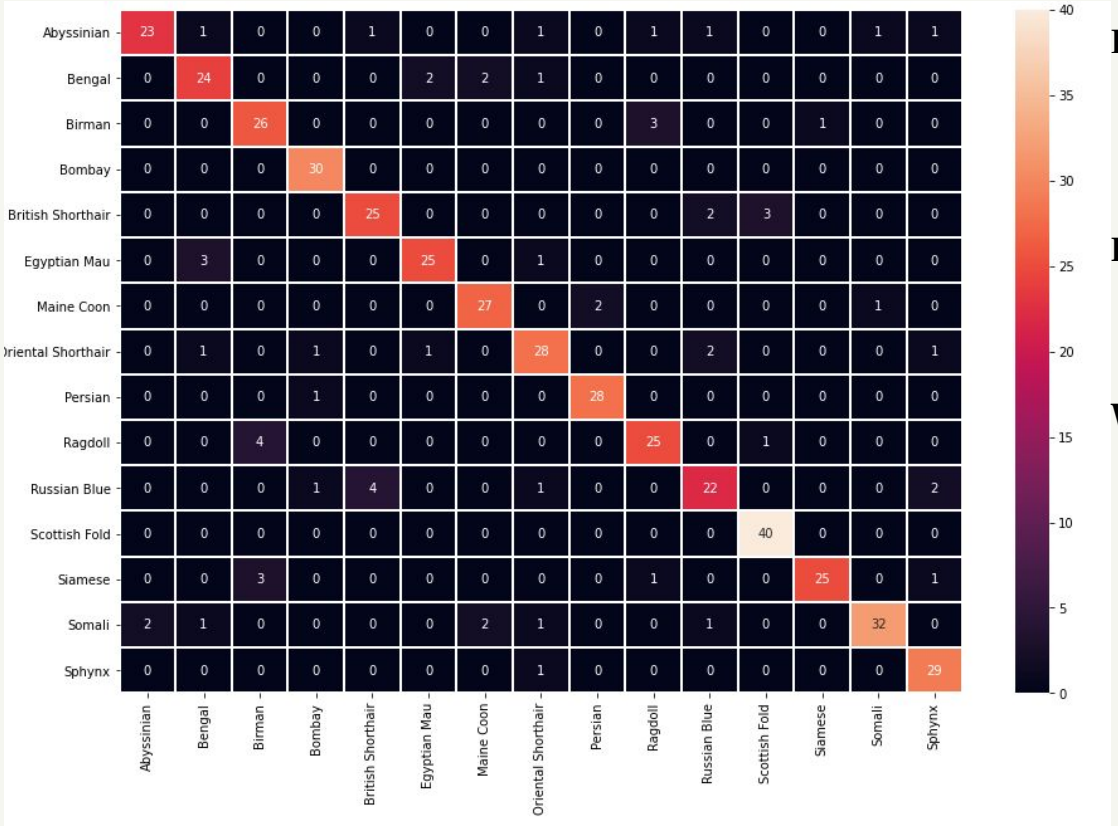
BIG TRANSFER (BiT)

What is Big Transfer?

- Adaptation of ResNetv2-152
- State of the art performance FGIC tasks
- Replaces batch normalization with:
  - ***Group normalization***
  - ***Weight standardization***
- Why?
  - BN's error increases rapidly when the batch size becomes smaller
  - GN can be naturally transferred from pre-training to fine-tuning
  - GN significantly outperforms batch normalization for transfer learning tasks



BIG TRANSFER (BiT) MODEL



DETAILS |

- Optimizer: **SGD**
- Loss function: **Cross-entropy**
- LR: `5.2e-2`

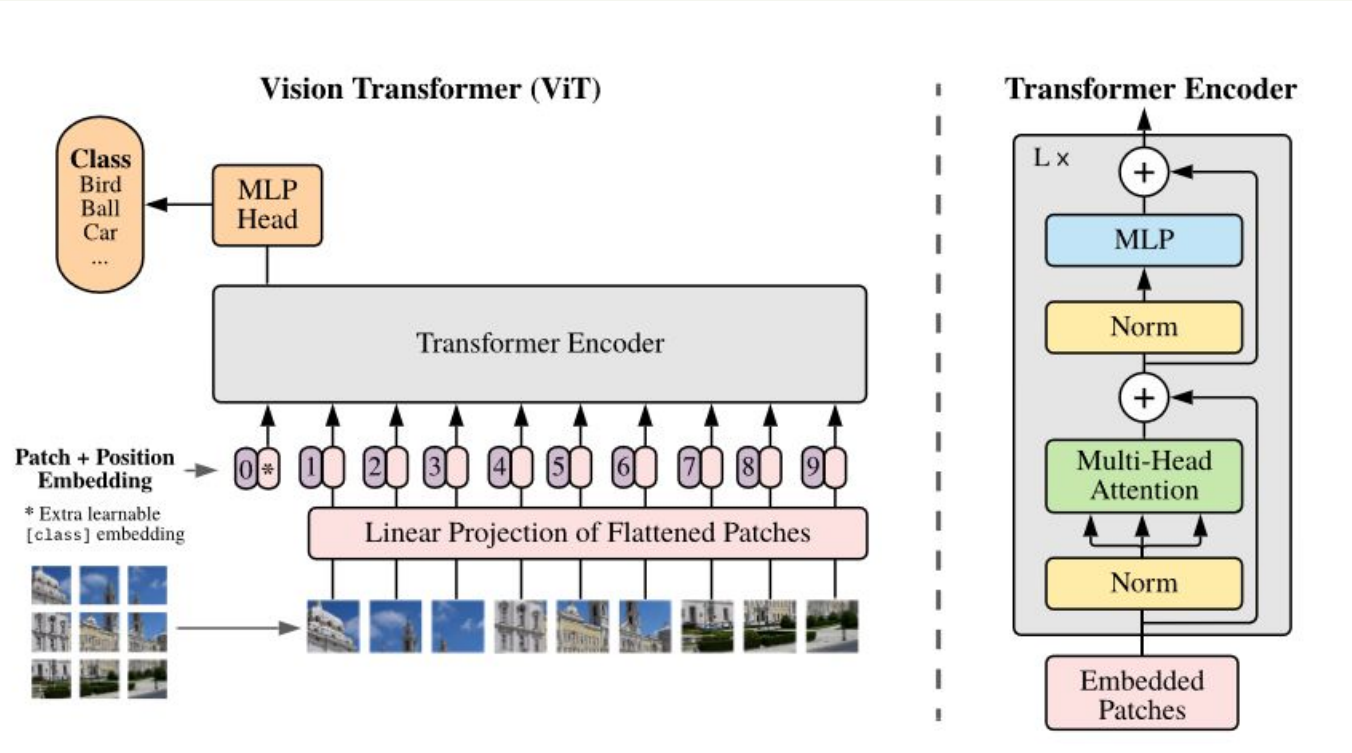
RESULTS |

- Validation Accuracy: **91%**
- Test Accuracy: **87%**

WHAT CATS IS IT MISSING? |

- Birman vs. Ragdoll
- Bengal vs. Egyptian Mau
- Russian Blue vs. British Shorthair
- These, it turns out, are persistent problem breeds throughout the data…. hmmm……
    - We'll come back to this

VISION TRANSFORMERS



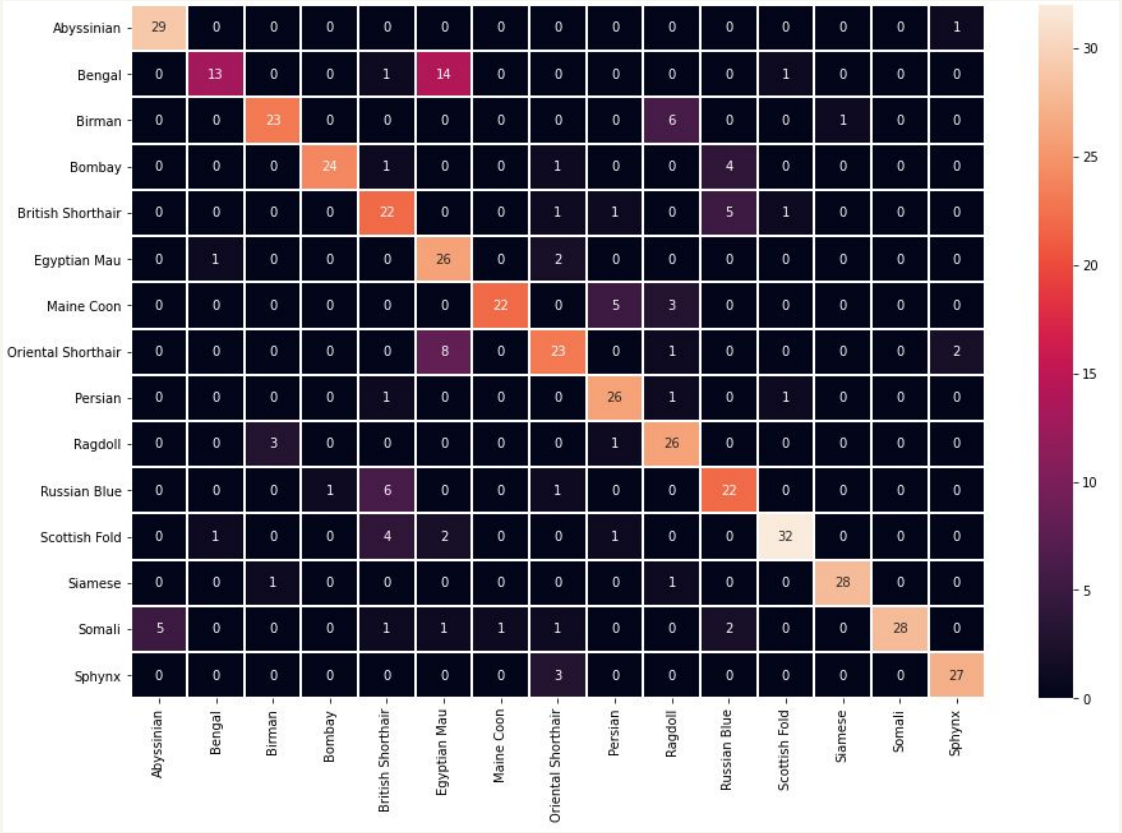
A Review From Earlier!

- Adapted from popular NLP methods
- SPLITS image into fixed-size regions
- LINEARLY PROJECTS the regions which...
- ENCODES information about the POSITION of features.
- Can be blended with CNNs into a hybrid model - not much work has been done here, though.

Vision Transformer performs better with high resolution images.

- Major impact on the performance of my model!

VISION TRANSFORMER MODEL



DETAILS |

- Optimizer: **SGD**
- Loss function: **Cross-entropy**
- LR: `0.001`

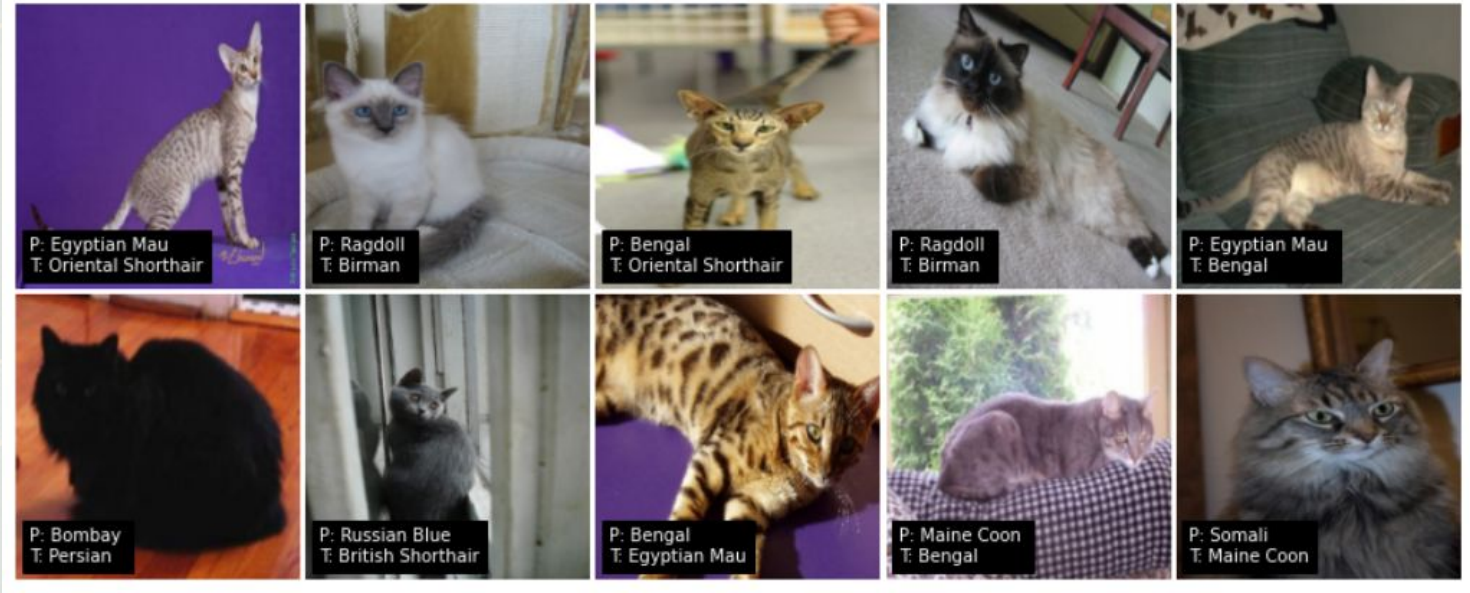
RESULTS |

- Validation Accuracy: **85.85%**
- Test Accuracy: **78.94%%**

WHAT CATS IS IT MISSING? |

- Ragdoll v. Birman and Russian Blue vs. British Shorthair are still causing problems...
- ***But what's up with the Bengal???***

SO, THE DATA HAS ISSUES



SO, THE DATA HAS ISSUES

THE DATA ARE OFTEN WRONG

These are not bengals! The cat on the left appears in the data 3 times

THE DATA ARE OFTEN WRONG


These are not bengals! The cat on the left appears in the data 3 times


These are not even domestic cats!!!

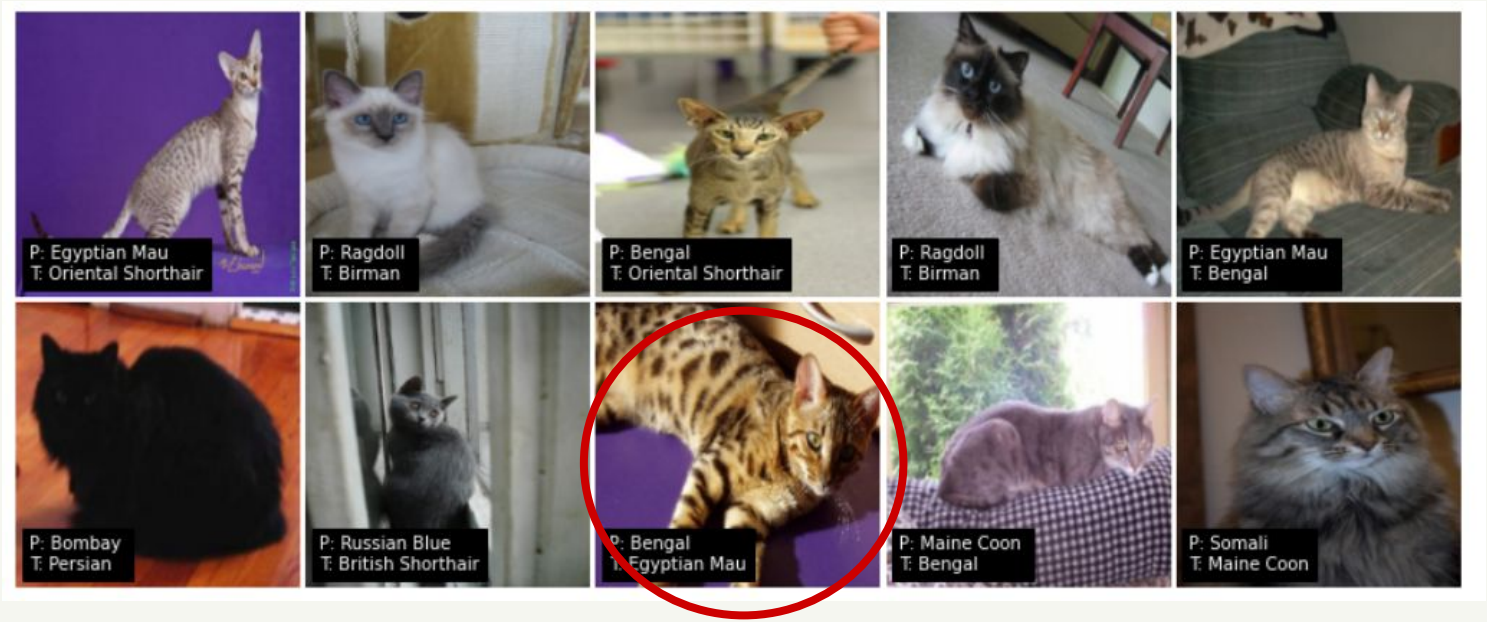
THE DATA ARE OFTEN WRONG


These are not bengals! The cat on the left appears in the data 3 times


These are not even domestic cats!!!

And some state of the art models are getting 97% test accuracy? I have questions.

THE MODEL WAS RIGHT!



Notice the rosette spots? This is a Bengal!

CONCLUSION

- Pre-trained deep neural networks have good performance on fine-grained image classification tasks
- Vision transformers and self-attention models offer a promising and transformative future for the FGIC
    - Provided decent test accuracy even with messy, low-quality data
- Take accuracy results with a grain of salt…
    - I would be more concerned if the model were getting 95%+ test accuracy than 80-90%.
    - Garbage in, garbage out

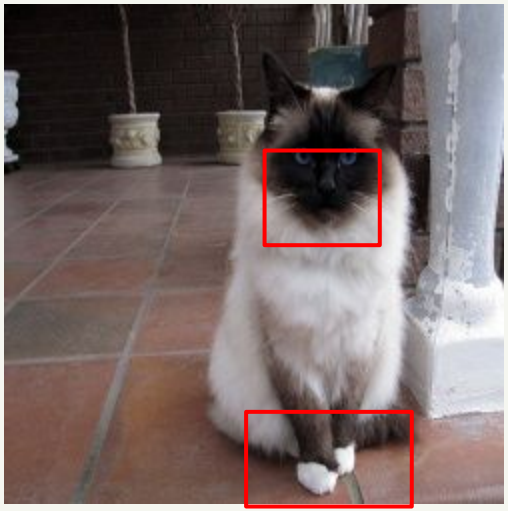
QUESTIONS

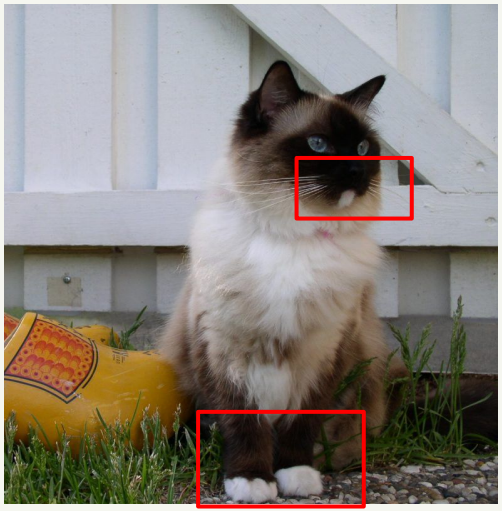
SELECTED REFERENCES

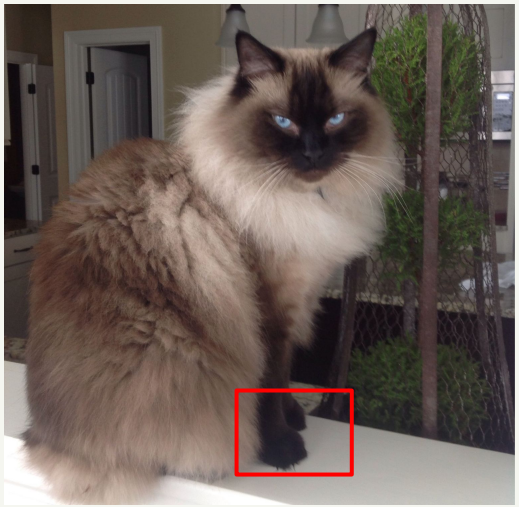
- Dostovitskiy, A., et al., An Image is Worth 16 x 16 Words: Transformers for Image Recognition at Scale. aRXiv preprint. https://arxiv.org/pdf/2010.11929v1.pdf (2020)
- Kolesnikov, A., Beyer, L., Xiaohua, Z., Puigcerver, J., Yung, J., Sylvain, G., Houlsby, N. Bit Transfer (BiT): General Visual Representation Learning. arXiv preprint https://arxiv.org/pdf/1912.11370v3.pdf (2020)
- Parkhi, A., Vedaldi, A., Zisserman, C., & Jawahar, V. Cats and Dogs. IEEE Conference on Computer Vision and Pattern Recognition. (2012)
- Peng, Y., He, X., & Zhao, J., Object-Part Attention Model for Fine-Grained Image Classification. IEEE Transactions on Image Processing, 27, 3. 10.1109/TIP.2017.2774041 (2018)
- Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens.Stand-alone self-attention in vision models. InNeurIPS, 2019
- Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Weight standardization. arXivpreprint arXiv:1903.10520, 2019.
- Wei, X., Wu, J., Quan, C. Deep learning for fine-grained image analysis: A survey. aRXiv preprint. https://arxiv.org/pdf/1907.03069.pdf (2019)
- Wu, Y., & He, K. Group Normalization. arXiv preprint. https://arxiv.org/abs/1803.08494 (2018)

THE BIRMAN vs RAGDOLL PROBLEM

Unless you have a full image, you cannot always tell the breeds apart. Much of the training data for Birmans does not show their feet. So what is the model learning to detect?







Birman

Ragdolls

THE BIRMAN vs RAGDOLL PROBLEM

Also, almost all of the data for Ragdolls depicts bicolor Ragdolls - if the data is not trained on mitted Ragdolls (white chim, white paws) how can we expect the model to classify them correctly?



Ambiguous Birman training data. Are these even Birmans? I don't know.



Ambiguous Ragdoll training data. Some errors are to be expected.

INTERNAL + EXTERNAL CONSISTENCY

This Birman kitten appeared in the test data twice and was consistently (across both observations for both BiT and ViT) classified as a Ragdoll. The kitten has a lighter coloration and a lighter chin than many Birmans in the training data. This may help us parse out the schematic features the model is picking up on.