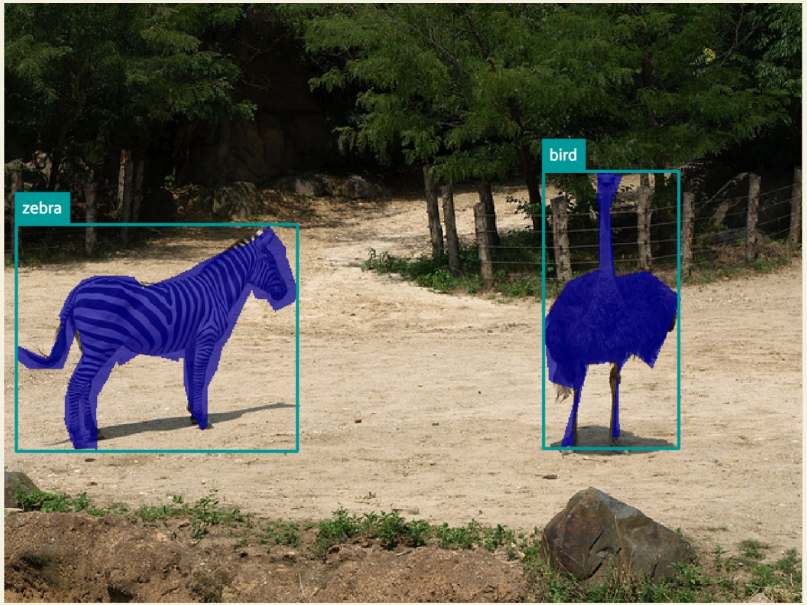
Classifying Animal Images

Matt Zuk

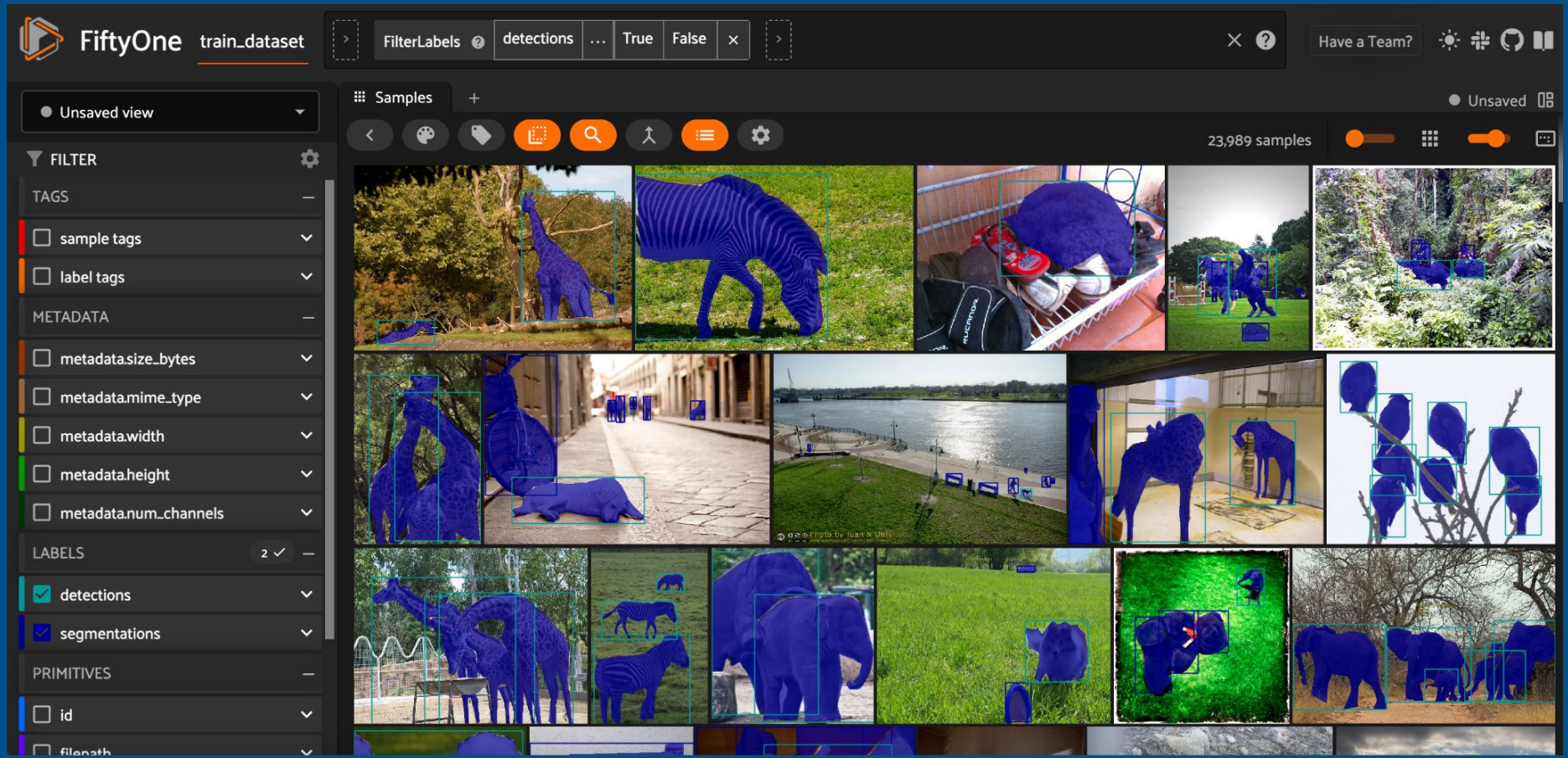


Overview of Project/Objectives

- Classifying images of animals
- Generating captions based on these classifications
- Employ CNNs and transformers
- Combine computer vision with natural language processing
    - Range of applications beyond this project

About the Data

- Original data: 2017 COCO dataset
  - 118,287 images originally in training data
  - Each with its own annotations
- Data Cleaning
  - Used 23,989 images containing animals
  - Resized and normalized
- FiftyOne library
- 2014 COCO dataset used for testing later on


FiftyOne
train_dataset
FilterLabels
detections
...
True
False
Have a Team?
Unsaved view
FILTER
TAGS
sample tags
label tags
METADATA
metadata.size_bytes
metadata.mime_type
metadata.width
metadata.height
metadata.num_channels
LABELS
2
detections
segmentations
PRIMITIVES
id
filepath
Samples
Unsaved
23,989 samples

Part One: CNN to Classify Animal Images

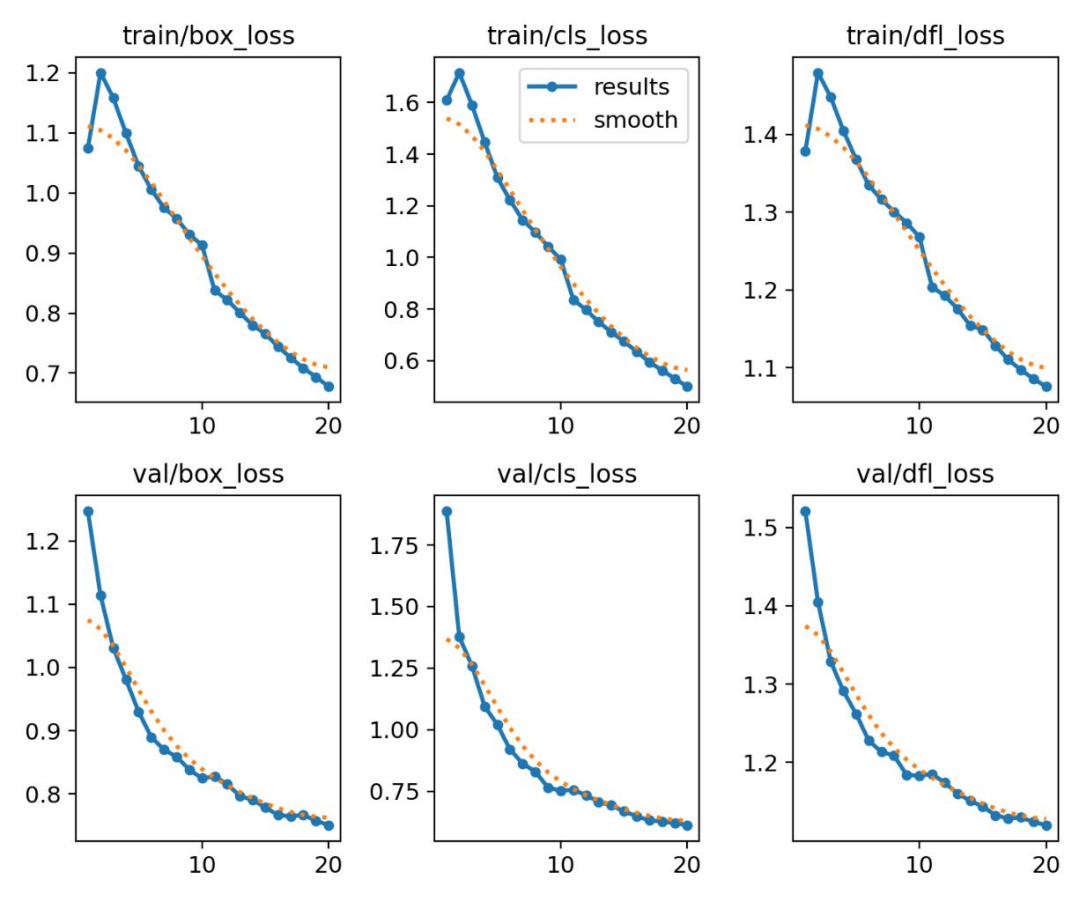
ML Model Used: YOLO

- YOLOv8

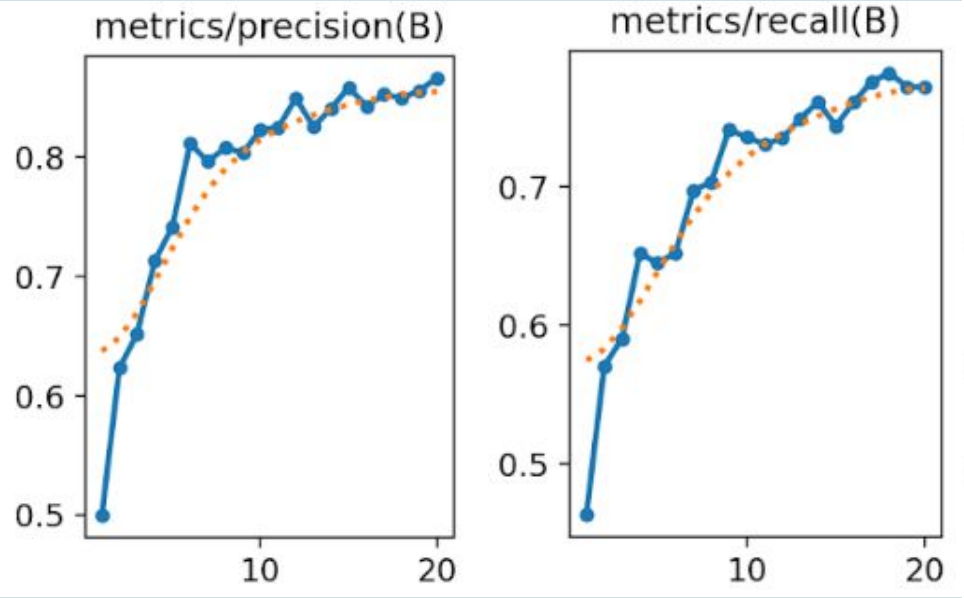
Two different approaches:

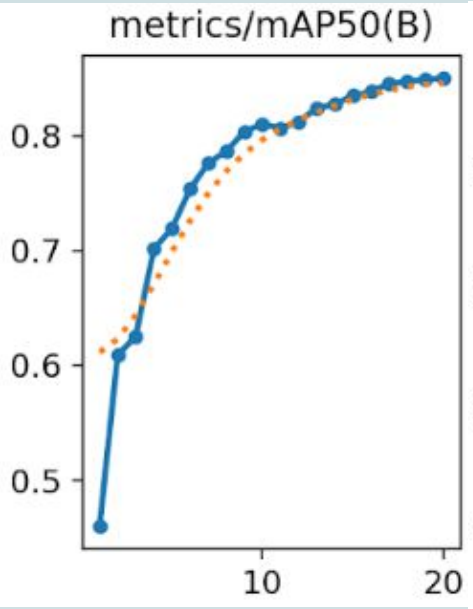
- To start:
  - Using CPU, second smallest variant of YOLO
- ★ Second approach:
  - Using GPU, largest variant of YOLO
  - 20 epochs

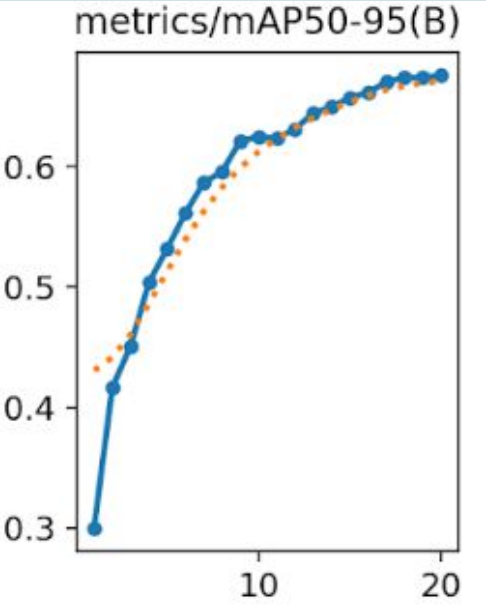
Train & Validation Loss Graphs



Performance Metrics (Train/Val)





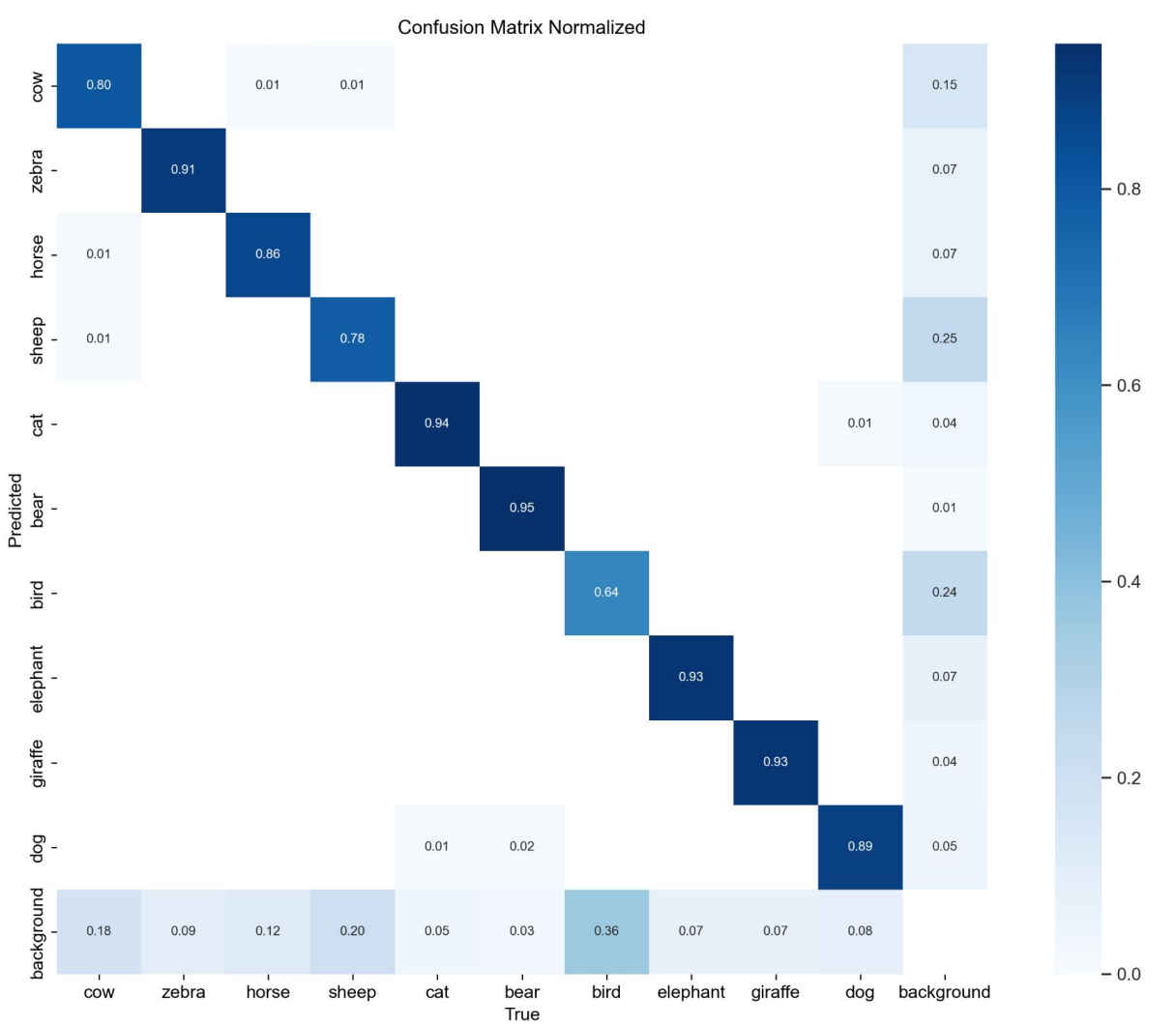


In attempt to improve the model, we tried to train it even more → overfitting! Returned back to original model.

Test Results

Class	Images	Instances	Precision	Recall	mAP@50	mAP@50-95
all	8265	21806	0.903	0.833	0.904	0.754
cow	666	2841	0.877	0.764	0.867	0.68
zebra	677	1886	0.923	0.878	0.95	0.808
horse	1001	2194	0.931	0.834	0.918	0.753
sheep	489	3216	0.817	0.722	0.814	0.623
cat	1480	1669	0.941	0.934	0.968	0.846
bear	341	462	0.936	0.944	0.971	0.875
bird	1121	3956	0.829	0.569	0.702	0.49
elephant	714	1863	0.906	0.899	0.946	0.812
giraffe	849	1767	0.945	0.908	0.958	0.837
dog	1521	1952	0.924	0.877	0.941	0.813

Confusion Matrix




COCO_val2014_000000505733.jpg
bird 0.9
bird 0.9
giraffe 0.9
COCO_val2014_000000410097.jpg
sheep 0.9
sheep 0.9
sheep 0.5
COCO_val2014_000000211204.jpg
bear 0.9
COCO_val2014_000000
cow 0.5
giraffe 0.9
giraffe 0.9
giraffe 0.8
COCO_val2014_000000125412.jpg
bear 0.3
bear 0.9
bear 0.9
COCO_val2014_000000105734.jpg
bird 0.8
bird 0.9
bird 0.3
bird 0.8
bird 0.7
cat 0.5
cat 0.9
COCO_val2014_000000131326.jpg
elephant 0.9
elephant 0.8
elephant 0.9
elephant 0.5
sheep 0.5
sheep 0.3
sheep 0.3
sheep 0.6
cat 0.9
elephant 1.0
COCO_val2014_000000054627.jpg
horse 0.9
horse 0.5
COCO_val2014_000000521593.jpg
zebra 0.3
zebra 0.8
giraffe 0.8
zebra 0.8
zebra 0.3
COCO_val2014_000000286327.jpg
sheep 0.9
dog 0.9

Part Two: Caption Generating

ML Model Used for Generating Captions

- BlipForConditionalGeneration and BlipProcessor
- Vision Language Model

Two different approaches:

- To start:
  - Feeding only images into Blip
- ★ Second approach:
  - Feed images into Blip
  - Also feed in filtered output from YOLO model

Examples

YOLO + Blip Generated Caption: 4 sheep in a pen

Blip Generated Caption: the sheep are white

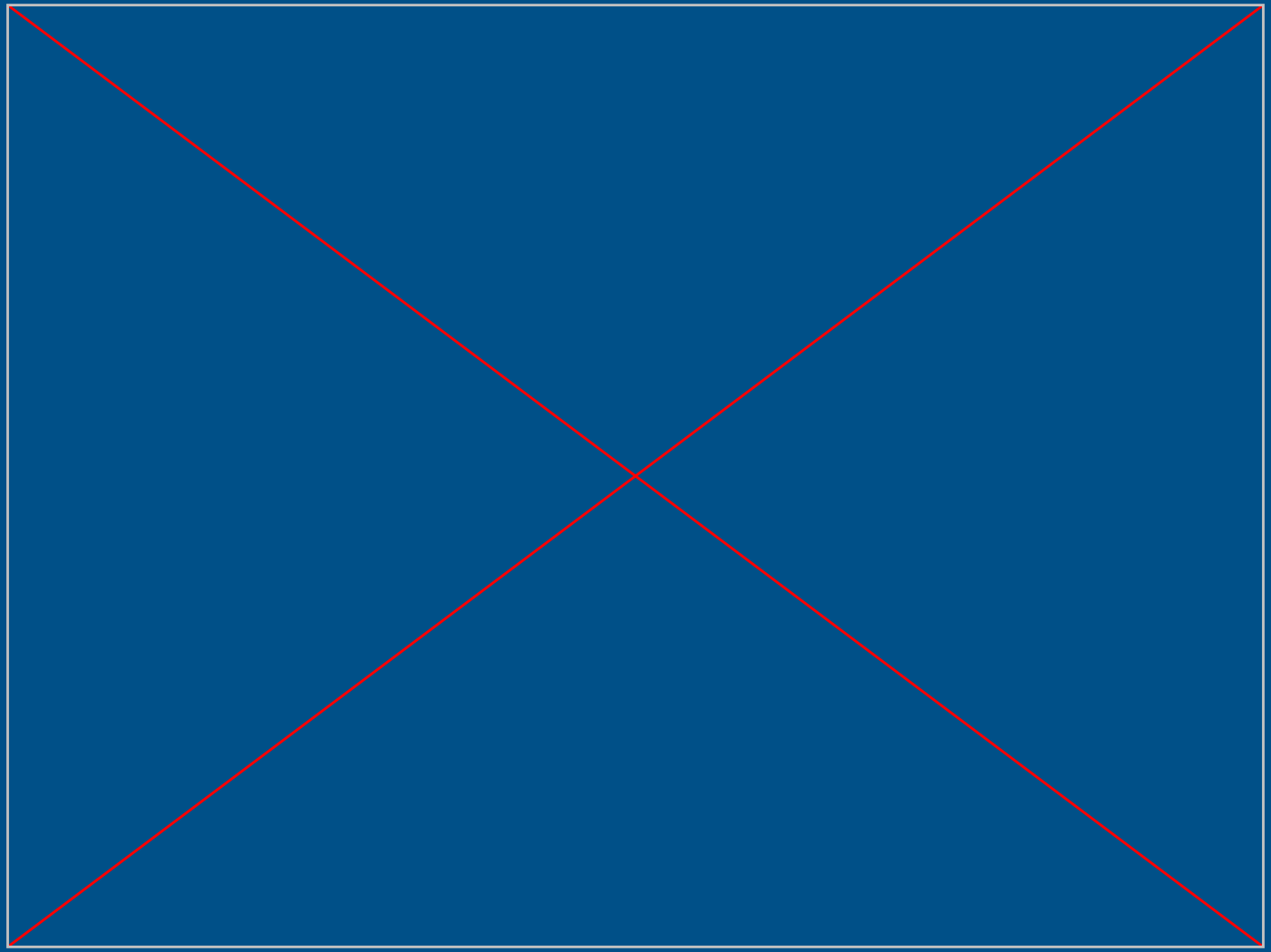
YOLO + Blip Generated Caption: dogs are sitting in a col of photos

Blip Generated Caption: a col of dogs

YOLO + Blip Generated Caption: horses are standing in a circle

Blip Generated Caption: a group of people dressed in medieval costumes

Demo



Limitations

- Amount of data → add more data
- Amount of animals trained on → train on other types of animals
- Hardware/Lack of Processing Power - Blip Model
- Overlap between 2014 and 2017 Dataset

Areas for Improvement/Next Steps

- Investigate what is strongly contributing to errors
  - Background noise vs overlapping features
- Reduce misclassifications
  - Augment the training dataset
  - Balance the dataset
  - Class-specific loss tuning