Classification of Sketch for Information Retrieval

Final Presentation

Akshit Kumar (EE14B127)
Sachin Agrawal (EE14B104)
Advisor : Mahesh Mohan

Indian Institute of Technology, Madras

Table of contents

Introduction

How is our problem different from Normal Image classification?

How is our problem different from Normal Image classification?

- Sketches are visually less complex than photographs

How is our problem different from Normal Image classification?

- Sketches are visually less complex than photographs
    - Photographs have 3 channels,

How is our problem different from Normal Image classification?

- Sketches are visually less complex than photographs
  - Photographs have 3 channels,whereas Sketches are encoded as black-and-white

How is our problem different from Normal Image classification?

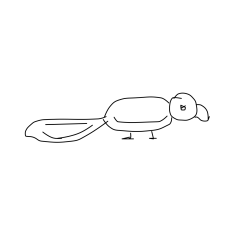
- Sketches are visually less complex than photographs
    - Photographs have 3 channels,whereas Sketches are encoded as black-and-white
    - Photographs contain visual information throughout the image,

How is our problem different from Normal Image classification?

- Sketches are visually less complex than photographs
  - Photographs have 3 channels,whereas Sketches are encoded as black-and-white
  - Photographs contain visual information throughout the image, Sketches are sparse

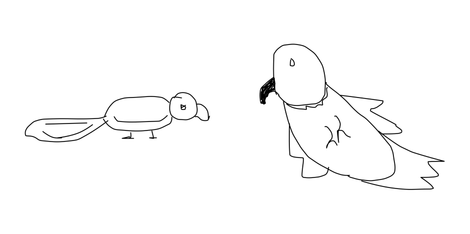
How is our problem different from Normal Image classification?

- Sketches are visually less complex than photographs
  - Photographs have 3 channels,whereas Sketches are encoded as black-and-white
  - Photographs contain visual information throughout the image, Sketches are sparse
- Sketches and Photographs have different sources of intraclass variation
  - Sketches vary widely based on artistic style



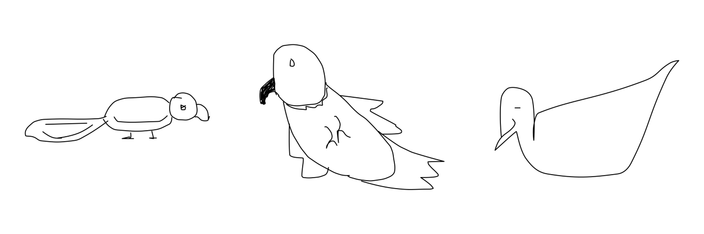
How is our problem different from Normal Image classification?

- Sketches are visually less complex than photographs
  - Photographs have 3 channels,whereas Sketches are encoded as black-and-white
  - Photographs contain visual information throughout the image, Sketches are sparse
- Sketches and Photographs have different sources of intraclass variation
  - Sketches vary widely based on artistic style



How is our problem different from Normal Image classification?

- Sketches are visually less complex than photographs
  - Photographs have 3 channels,whereas Sketches are encoded as black-and-white
  - Photographs contain visual information throughout the image, Sketches are sparse
- Sketches and Photographs have different sources of intraclass variation
  - Sketches vary widely based on artistic style



Applications

- A search engine add-on (Aid Google's Image search)

Applications

- A search engine add-on (Aid Google's Image search)
- Pictionary: An AI to play pictionary
- Future of human computer interfaces

Applications

- A search engine add-on (Aid Google's Image search)
- Pictionary: An AI to play pictionary
- Future of human computer interfaces
- Educational tools

Dataset and Data Augmentation

Dataset and Data Augmentation

- TU Berlin dataset collected using crowd-sourcing of sketches.
- Dataset has 250 classes, and 80 images per class
- Reduce the original image size of 1111×1111 pixels to 128×128 pixels using bilinear interpolation
- Data augmentation using flipped images
- Rotating original and flipped images by $5°, 10°$ and $15°$
- Hence, we obtain 960 images per class

Past Work

Past Work

- The first attempts at classifying sketches made use of hand crafted features and SVM classifiers. They were able to achieve accuracy of 56% -Eitz et al.(SIGGRAPH 2012)

Past Work

- The first attempts at classifying sketches made use of hand crafted features and SVM classifiers. They were able to achieve accuracy of 56% -Eitz et al.(SIGGRAPH 2012)
- Sketch-a-Net that beats humans by Qian et al., 2015 makes use of stroke sequentiality to convert a sketch into a multiple channel image. They also make use of wide filters for learning sparse features. This approach reaches an accuracy of 74.9%.

Past Work

- The first attempts at classifying sketches made use of hand crafted features and SVM classifiers. They were able to achieve accuracy of 56% -Eitz et al.(SIGGRAPH 2012)
- Sketch-a-Net that beats humans by Qian et al., 2015 makes use of stroke sequentiality to convert a sketch into a multiple channel image. They also make use of wide filters for learning sparse features. This approach reaches an accuracy of 74.9%.
- DeepSketch 2 - Seddati et. al, 2016 generates partial sketches from the given dataset. Partial sketches are used to fine-tune the network. The final accuracy achieved is 77%

Vanilla CNN - Sketch-A-Net

- We start with training a vanilla CNN for image classification
- We use wide filters in the first layer
- This helps us in capturing the sparse features in the sketch
- Architecture inspired from Sketch-a-Net
- Implementation in Tensorflow
- We obtain an accuracy of 62% with this approach
- This is better than the 56% presented by Eitz et. al in the initial paper

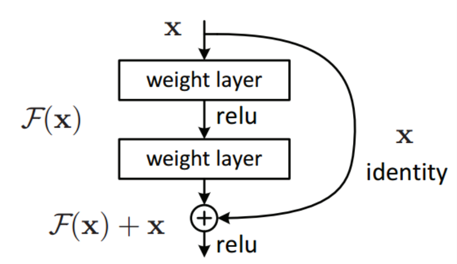
Vanilla CNN Architecture

Layer	Type	Filter size	Filter Num	Stride	Output Size
L1	Input	-	-	-	225×225
	Conv	15×15	64	3	71×71
	ReLU	-	-	-	71×71
	MaxPool	3×3	-	2	35×35
L2	Conv	5×5	128	1	31×31
	ReLU	-	-	-	31×31
	MaxPool	3×3	-	2	15×15
L3	Conv	3×3	256	1	15×15
	ReLU	-	-	-	15×15
L4	Conv	3×3	256	1	15×15
	ReLU	-	-	-	15×15
L5	Conv	3×3	256	1	15×15
	ReLU	-	-	-	15×15
	MaxPool	3×3	-	2	7×7
L6	Conv	1×1	512	1	1×1
	ReLU	-	-	-	1×1
	Dropout(0.50)	-	-	-	1×1
L7	Conv	1×1	512	1	1×1
	ReLU	-	-	-	1×1
	Dropout(0.50)	-	-	-	1×1
L8	Conv	1×1	250	1	1×1

ResNet-Training and results

Training on ResNet

- State of the art for image classification
- Use wide filters to learn sparse features in the image
- Current architecture consists of 25 layers



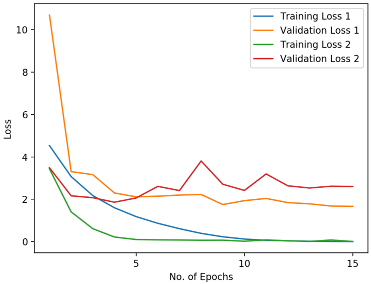
Training on ResNet - Architecture

Layer	Output Size
Input	$128 \times 128 \times 1$
7×7 conv, 64, /2	$64 \times 64 \times 64$
3×3 residual unit, 64	$64 \times 64 \times 64$
3×3 residual unit, 64	
3×3 residual unit, 64	
3×3 residual unit, 128, /2	$32 \times 32 \times 128$
3×3 residual unit, 128	
3×3 residual unit, 128	
3×3 residual unit, 256, /2	$16 \times 16 \times 256$
3×3 residual unit, 256	
3×3 residual unit, 256	
3×3 residual unit, 512, /2	$8 \times 8 \times 512$
3×3 residual unit, 512	
3×3 residual unit, 512	
8×8 Average Pooling	512
Fully Connected, 250	250

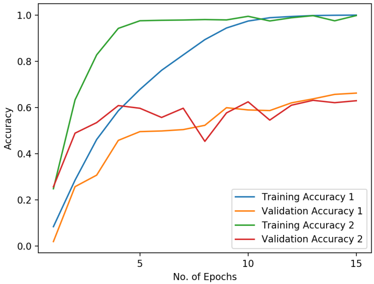
Training on ResNet - Details

	w/o Data Augmentation	w/ Data Augmentation
Training Time(CPU)	N.A	$\sim$ 3 days
Training Time(GPU)	$\sim$ 2.5 hours	$\sim$ 7.5 hours
Learning Rate	$1e-3$	$1e-3$
Training Images	24000 ($48 \times 2 \times 250$)	72000 ($48 \times 6 \times 250$)
Validation Images	4000 (16×250)	4000 (16×250)
Test Images	4000 (16×250)	4000 (16×250)
Test Accuracy	66.2 %	62.9 %

Training on ResNet - Training Plots



(a) Loss Plots



(b) Accuracy Plots

Figure 1: Training and Validation Plots

Analysis of ResNet

- 3 Most Accurately Classified Classes

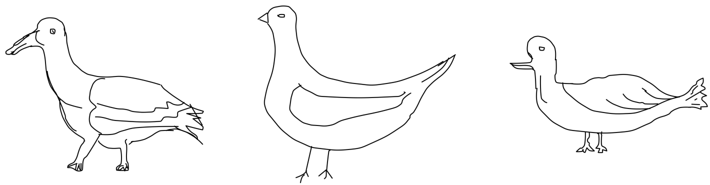
Rollerblades	1.000
Nose	1.000
Zebra	0.9375

- 3 Least Accurately Classified Classes

Dragon	0.1875
Seagull	0.125
Panda	0.0625

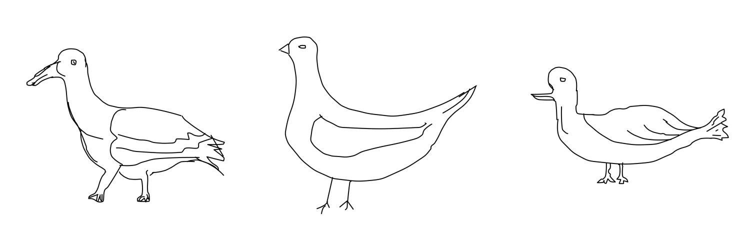
Possible Reasons for Misclassification

- Interclass Overlap (Eg. Seagulls)



Possible Reasons for Misclassification

- Interclass Overlap (Eg. Seagulls)

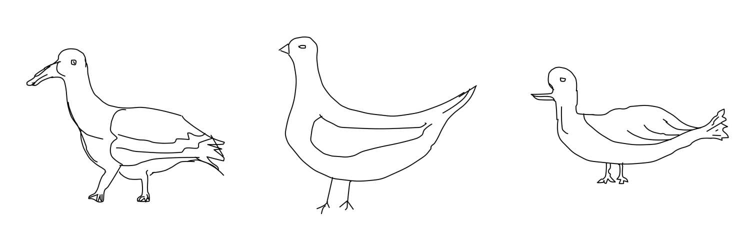


Seagull

Analysis of ResNet

Possible Reasons for Misclassification

- Interclass Overlap (Eg. Seagulls)

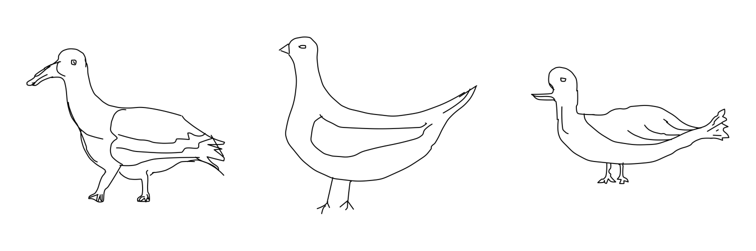


Seagull Pigeon

Analysis of ResNet

Possible Reasons for Misclassification

- Interclass Overlap (Eg. Seagulls)



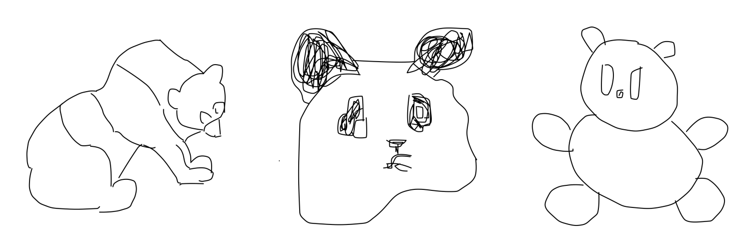
Seagull Pigeon Standing Bird

- No difference in drawing these varieties of birds
- Misclassified 5 times as pigeon and 3 times as standing bird

Analysis of ResNet

Possible Reasons for Misclassification

- Intraclass Variation and Interclass Overlap (Eg. Panda)



- Large variation in posing and color and artistic talent
- Classified as Teddy Bear one-third of the times

Training on ResNet - Limitation

- Test Accuracy below human performance ($\sim 73\%$)
- Training requires GPUs and is time expensive
- Cannot be used for light-weight applications such as search engine aid

Binary Sketch-A-Net

How to speed-up our network?

- XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks by Rastegari et. al.
- They proposed two networks.
- The first network made use of binary filters to perform convolution operations
- The second made use of binary filters as well as binarized the input to these filters
- We implement a binary sketch-a-net to optimize memory and computation efficiency

Binary Sketch-a-Net

- $58\times$ faster convolution
- $32\times$ memory saving
- We implement a binary sketch-a-net to optimize on computational efficiency

Binary Sketch-a-Net

$$I * W \approx (I \oplus B)\alpha$$

Where:

- I is the input
- W is the normal filter
- B is the filter consisting of $\{-1, 1\}$
- α is a positive number

Results from Binary Sketch-A-Net

- We obtained a top 1 accuracy level of 58.65%.
- We also obtain a top 5 accuracy level of 82.11%.
- This is lower than state of the art results.
- However, we achieved considerable memory savings and speed up.
- Our weights storage reduced from 32.7 MB to 1.82 MB. This is equivalent to $18\times$ memory savings.
- We also attained a $2\times$ speedup.

Summary

Architecture	Accuracy	Advantage
Vanilla Sketch-A-Net	62%	State of the Art Network
ResNet	66.2%	Improved Accuracy
Binary Sketch-A-Net	58.65%	Computational Efficiency

Conclusions and scope for future work

- We explored a ResNet, which gave us comparatively better accuracy
- We also explored a binary Sketch Net, which was computationally superior.
- One interesting experiment would be implementing a binary ResNet for sketch classification
- We were unable to get any interesting results from RNNs, but we still feel that this line of work needs to be explored further.

Thank You