SketchNet: Sketch Classification with Web Images[CVPR `16]

CS688 Paper Presentation 1

Doheon Lee

20183398

2018. 10. 23

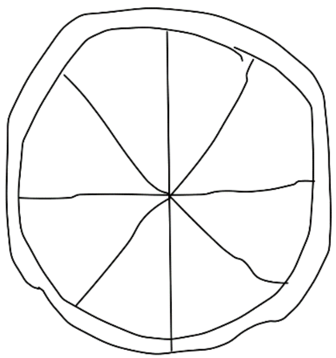
KAIST

Table of Contents

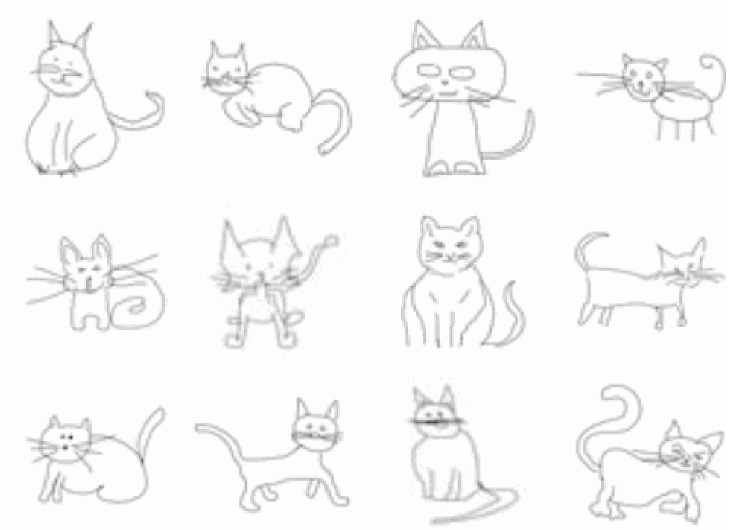
Introduction

Properties of Sketch Images

- Compared to Images
  - Texture less
  - Colorless
  - Different styles by people



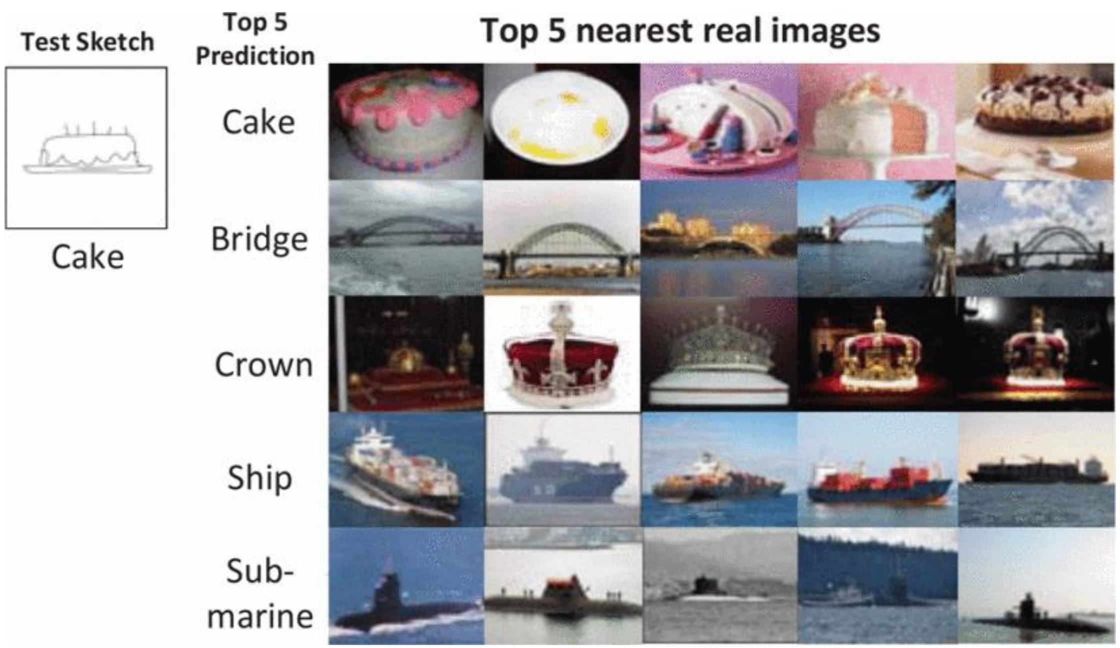
Pizza? Wheel?



Samples of cats drawn by human

Sketch-Based Image Retrieval

- **Find related image from sketch**
- **Large difference between sketch and image**

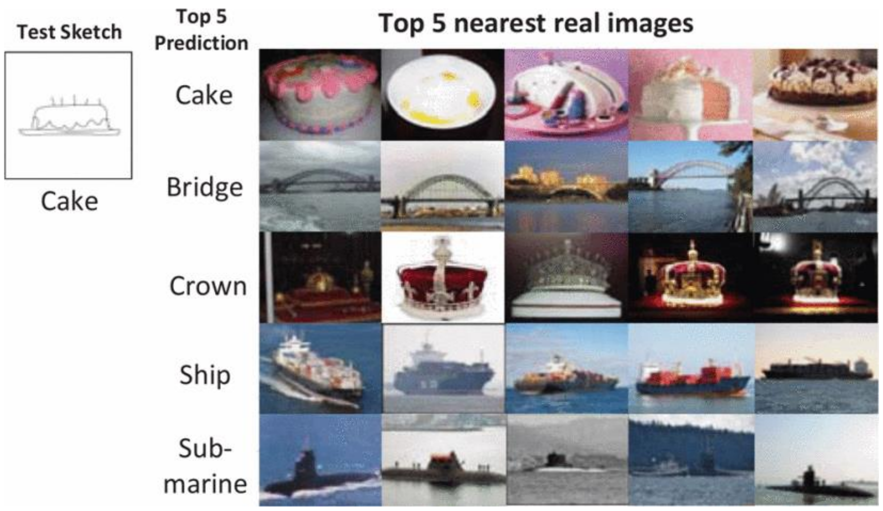


Relation between Image and sketch

- **Sketch is drawn from image**
- **Sketch-Based Image Retrieval can be considered as inverse task for drawing sketch**
- **Learn shared latent structures**

Inter class difference

- **Previous presentations are focus on intra-class difference**
- **This presentation work focuses on inter-class classification**

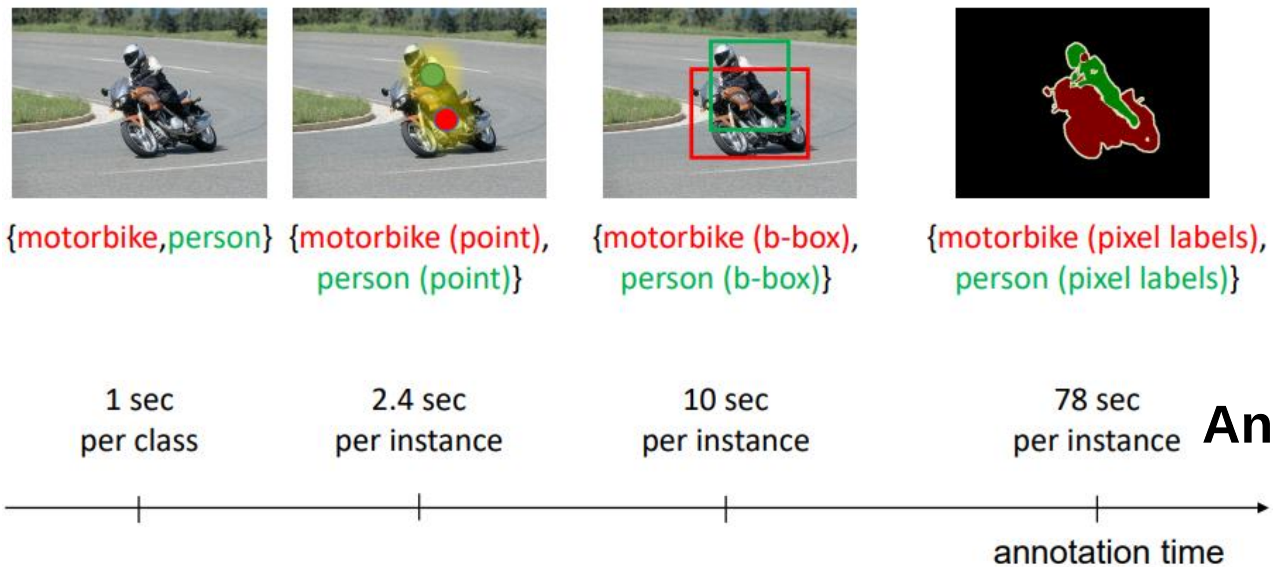


From chiwan's slide

Background

Manual Annotation

- **For supervised learning, we need a label for each datum**
- **However, high degree annotations are expensive**



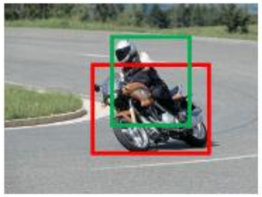
Manual Annotation time

Berman et al., What's the Point: Semantic Segmentation with Point Supervision, ECCV 16

KAIST

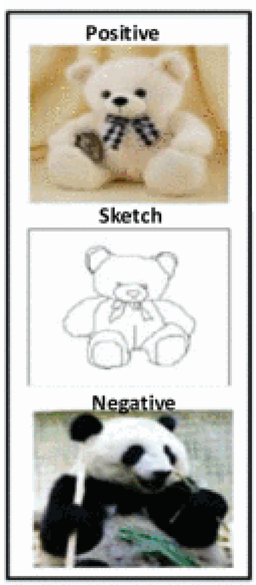
Weak Supervision

- **Lower degree annotation at train time than the required output at the test time**

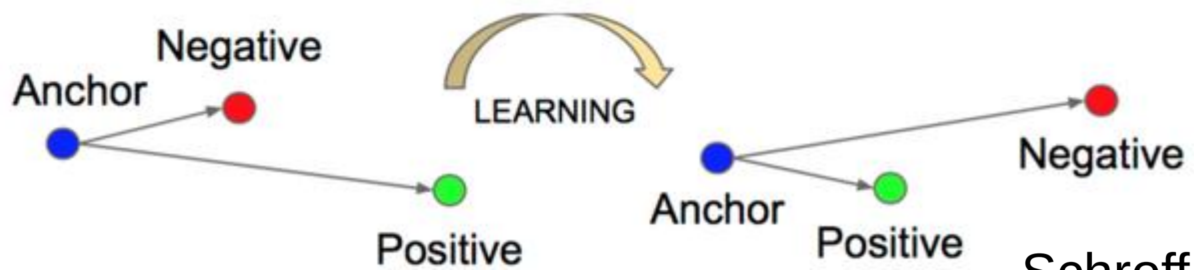
	Training Data	Target Data
(Regular) Supervised Learning	 {motorbike,person}	 {motorbike,person}
Weakly Supervised Learning	 {motorbike,person}	 {motorbike (b-box), person (b-box)}

Triplet Pair

- **Construct pair with positive and negative samples**
  - **Positive: similar image to anchor**
  - **Negative: Different image to anchor**



(a) Triplet Input



Schroff *et al*.

Make positive distance small, while negative difference large

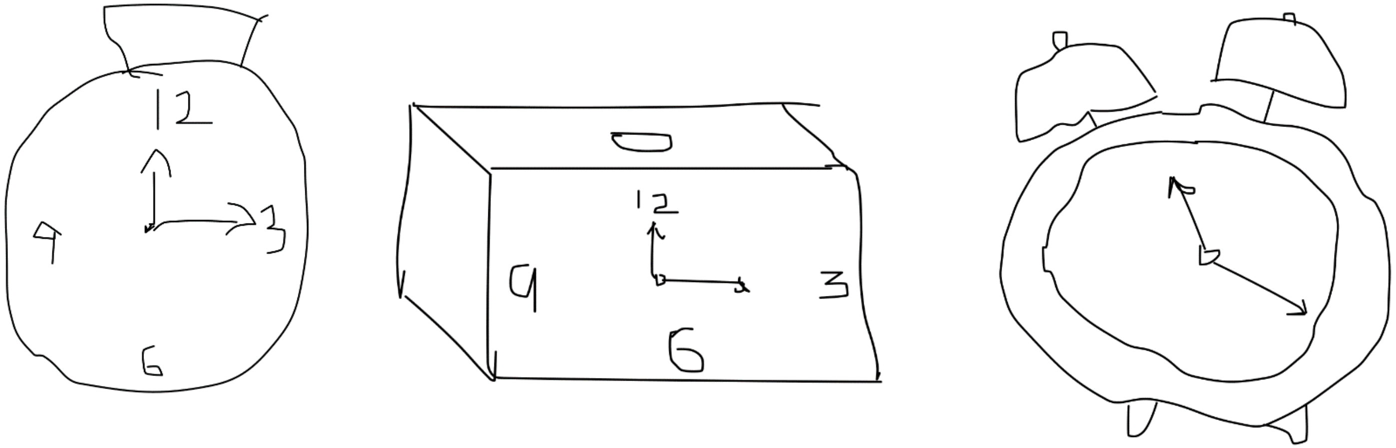
How Do Human Sketch Objects[TOG `12]

- **Construct Sketch Dataset: TU-Berlin**
  - **250 category**
  - **20K sketches**

- **Sketch classification from bag-of-features related SIFT[Lowe '04]**
  - **Limited to specific class of sketch with small variations**
  - **Represent a sketch as a frequency histogram of visual words**

How Do Human Sketch Objects[TOG `12]

- **Contents of TU-Berlin Dataset**
  - **Data labeled as "alarm clock"**



  - **80 instances for each 250 category**

SketchNet

Key Idea

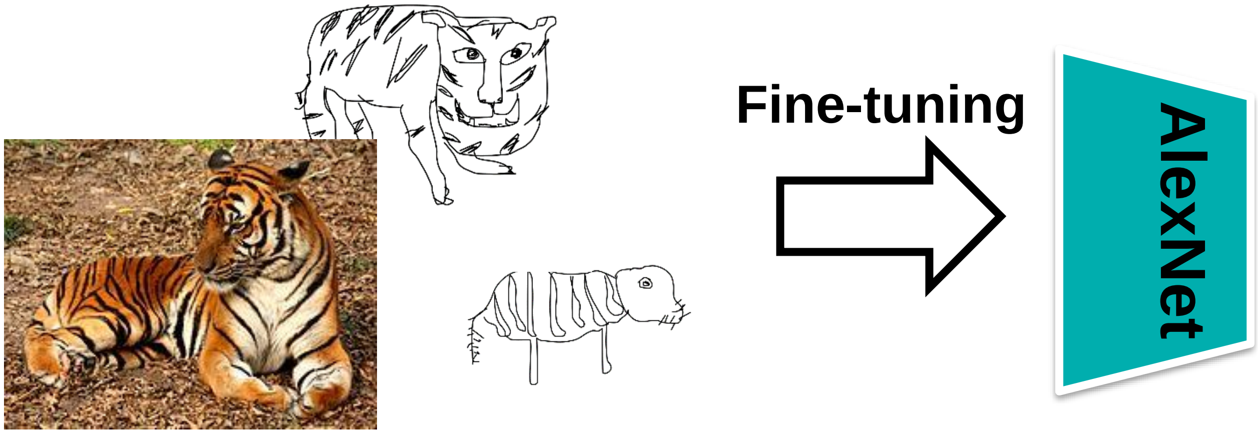
- **To Learn shared latent structures between sketch and image**

- **Construct triplet pair for sketch and images**

Construct training pair

- **Use Alexnet with pre-trained model on ImageNet**

- **Fine-tune with TU-Berlin dataset and collected Web Images**



Mixed dataset
(TU-Berlin and Web Images)

Construct training pair

- **For each sketch images, the nearest images in same category will have coherent appearance**

Find 5 nearest real images in “tiger” category

Sketch

…

“alarm clock”

…

“sun”

Find 5 nearest real images in each 5 wrong category

Find 5 most inaccurate categories

KAIST

Construct training pair

- **Now we have 5 positive images and 25 negative images**
  - **Construct 5x25 = 125 triplet pairs**



Sketch **Positive** **Negative**

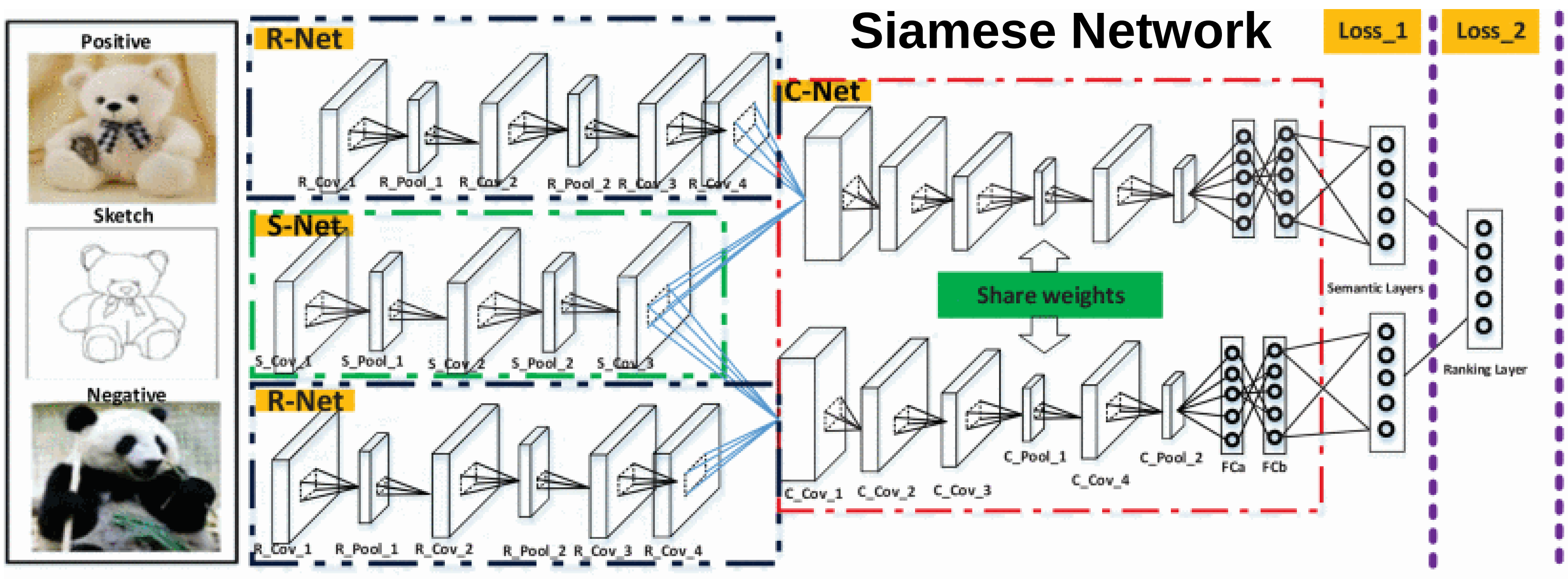
…



Sketch **Positive** **Negative**

Sketch Net network architecture

- **Because of significant gap between image and sketch, design new network**
  - **S-Net, R-Net, C-Net**



(a) Triplet Input (b) SketchNet architecture (c) Loss function

Sketch Net network architecture

- **S-Net: Learning sketch related features**
- **R-Net: Learning image related features**
- **C-Net: Merge feature maps between image and sketch**
  - **Make positive image pair generate higher score than negative image pair**

Loss function

- **Combine classification loss and ranking loss**
- **Classification loss**
  - **ability on image classification**

$$L_c(x^i, y^i, W_c) = -logP(y^i = k|x^i, W_c)$$
$$= -log\frac{e^{-f^k(x^i, W_c)}}{\sum_{l=1}^{C} e^{-f^l(x^i, W_c)}}$$

x: input image
y: input label
k: category label
W: weight
C: # of categories

- **Ranking loss**

$$L_r(\mathrm{p}_+^i, \mathrm{p}_-^i, y^i) = max(0, 1 - (\mathrm{p}_- - \mathrm{p}_+))$$

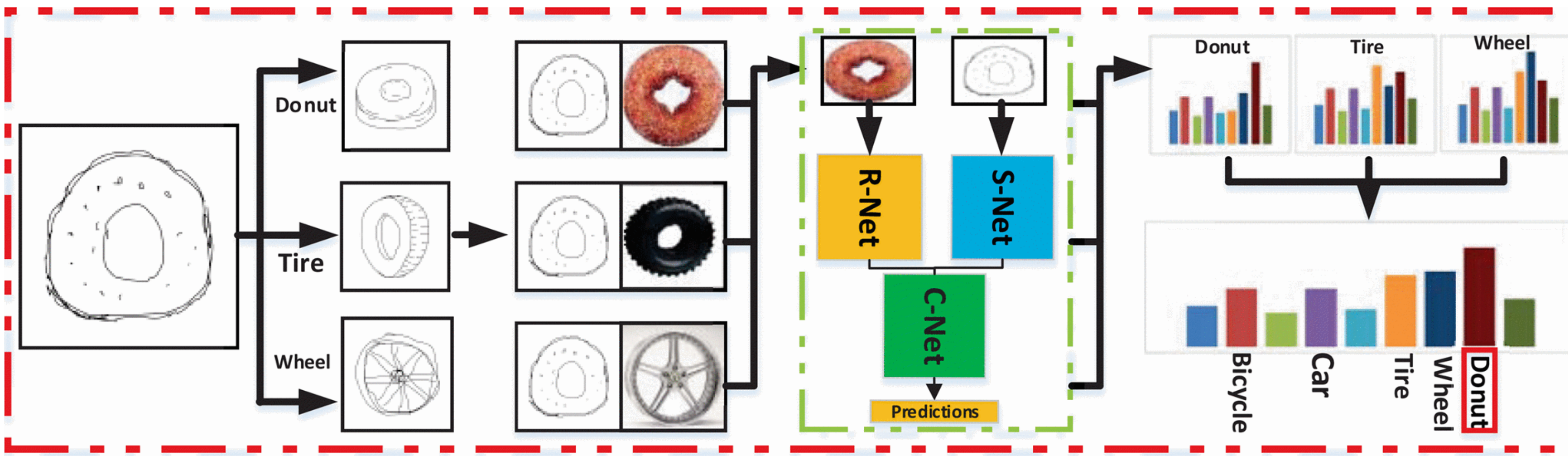
p+: positive pair score
p-: negative pair score

- **Loss function** $L_{SketchNet} = L_r + \lambda * Lc$

Testing Network

- **As we do not know label at the testing, triplet pair cannot be constructed**
  - **New network with**

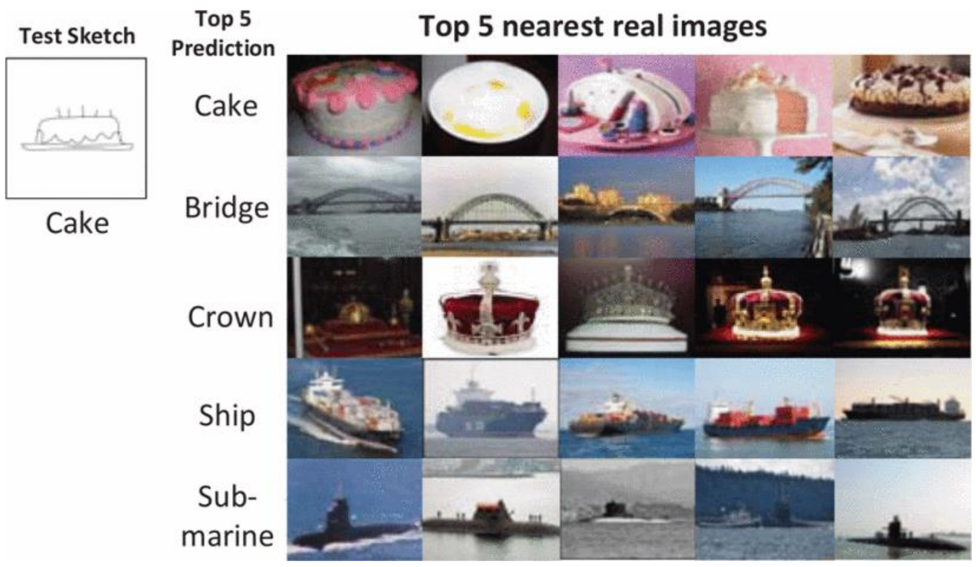
    One R-Net, S-Net and C-Net



KAIST

Testing Network

- For given sketch, using Alexnet, find 5 categories.
- For each category, find 5 nearest real images
- These image pairs are used for classification



Result

Experiment benchmark

- **The experiment are done in TU-Berlin dataset**
- **For each category, contains 80 data**
  - **The experiments are done in various test and training data ratio**

Experiment benchmark

# of training data

Table 2. The comparison classification results on TU-Berlin sketch benchmark

Methods	8	16	24	32	40	48	56	64	72
SketchNet	**58.04%**	**64.43%**	**67.89%**	**72.01%**	**73.54%**	**75.18%**	**76.08%**	**77.33%**	**80.42%**
SketchNet(no metric)	55.69%	64.37%	66.20%	71.19%	69.57%	73.62%	73.43%	76.50%	77.41%
AlexNet(mixed real images)	51.96%	59.22%	63.80%	65.97%	68.58%	69.80%	70.46%	72.31%	73.25%
AlexNet [20]	54.8%	62.3%	67.6%	68.12%	69.86%	71.65%	72.62%	74.02%	75.02%
GoogLeNet [32]	52.01%	59.61%	62.45%	67.48%	69.19%	70.5%	71.5%	72.4%	75.25%
NIN [23]	51.4%	61.9%	65.50%	68.05%	70.61%	71.50%	72.02%	73.82%	74.40%
VGGNet [5]	53.85%	60.65%	63.05%	65.54%	67.34%	69.54%	73.83%	75.17%	76.53%
FisherVector size 24 (SP) [29]	43%	52%	56%	59%	62%	65%	66%	67%	68%
FisherVector size 24 [29]	41%	50%	53%	56%	60%	62%	64%	64%	65%
FisherVector size 16 (SP) [29]	44%	50%	55%	57%	60%	63%	64%	65%	66%
FisherVector size 16 [29]	39%	45.5%	50%	53%	56%	59%	60%	61%	62%
Eitz et al. [12] (SVM soft)	33%	41%	44%	46%	50%	51%	54%	55%	55%
Eitz et al. [12] (SVM hard)	32%	37%	42%	45.5%	48%	49%	50.8%	53%	53%
Eitz et al. [12] (Knn soft)	26%	31%	34.8%	36%	39%	40.5%	42%	43%	44%
Eitz et al. [12] (knn hard)	22%	26%	28%	31%	33%	34.5%	35%	36%	37.5%

Thank you for Listening