Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval, ICCV '17

CS688 Paper Presentation

2018/10/23

20174315 Chiwan Song

KAIST

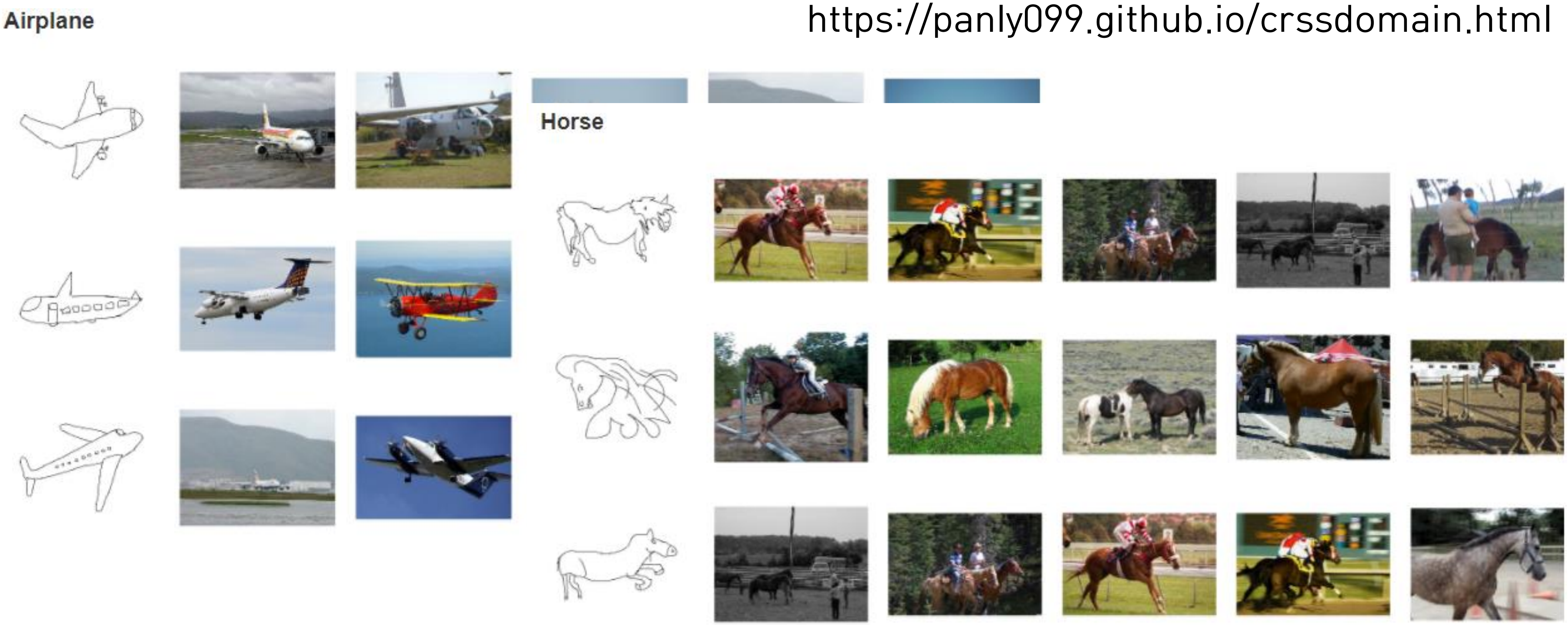
Table of contents

Introduction

Sketch-Based Image Retrieval

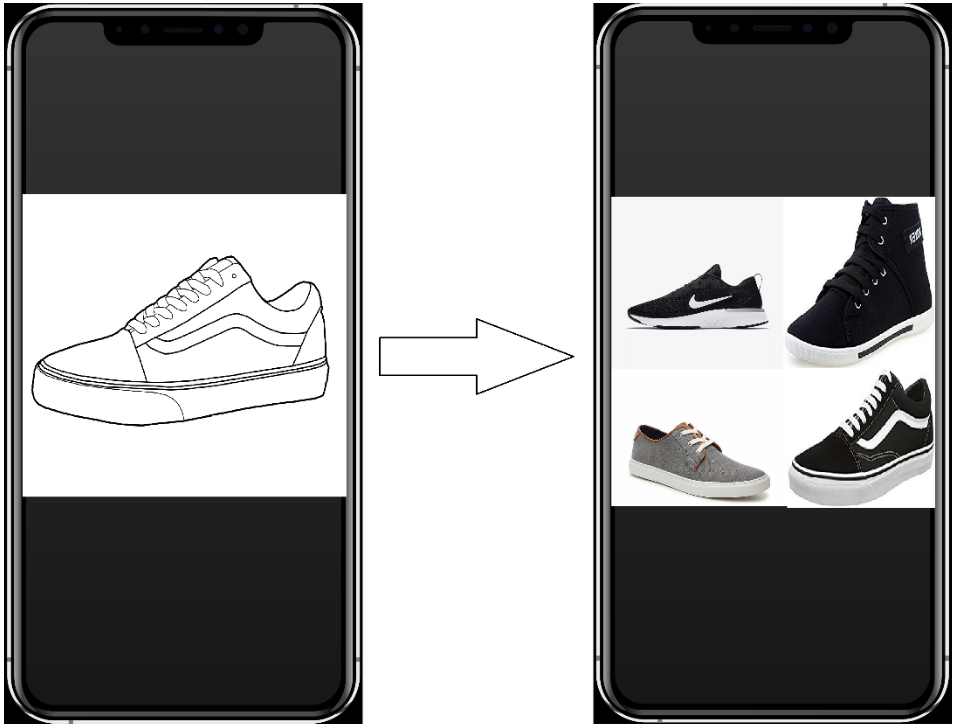
- It is a kind of **image retrieval!**
  - But, the queries are **freehand sketches**
- Examples

https://panly099.github.io/crssdomain.html



Sketch-Based Image Retrieval

- It can be applied to commercial applications
  - Searching online product catalogues for goods
  - By finger-sketching on a touch screen



Challenges in SBIR

- **Large domain gap** between sketch & photo

- Al
- Vi

Figure 1. FG-SBIR is challenging due to the misalignment of the domains (left) and subtle local appearance differences between a true match photo and a visually similar incorrect match (right).

ng

photo

 Song et al., ICCV '17

KAIST

Main contribution of the paper

- Deep Fine-grained SBIR model with **Attention & Coarse-to-Fine fusion**
  - Can keep both coarse and fine details to catch subtle difference between candidate photos
- New triplet loss function (HOLEF)
  - Make model **robust** against feature misalignment
- New dataset for SBIR

 Song et al., ICCV '17

Backgrounds

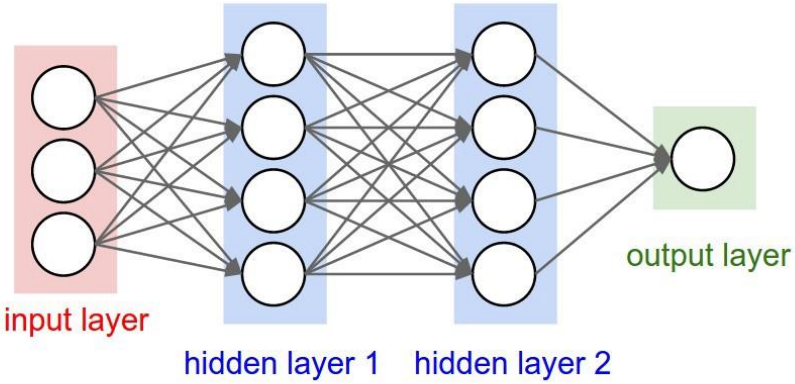
Backgrounds for understanding

- Properties of **Deep neural network**
- Concept of **Triplet loss**
- Concept of **Attention modeling**

 Song et al., ICCV '17

Deep Neural Network

- Inspired by **human neural system**
- Each layer has its own **weight and bias values**
  - $H(x) = Wx + b$
- Data is processed by each layer for purpose

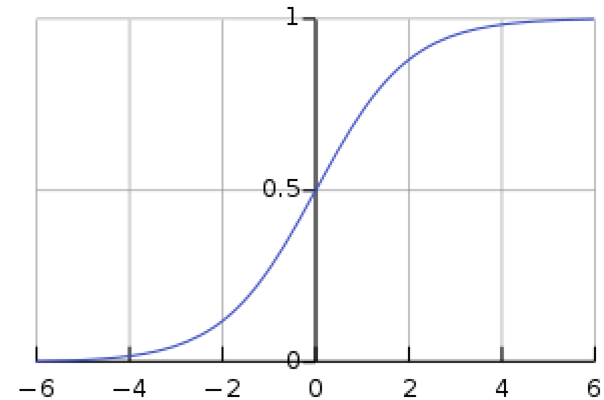


Simple deep neural network

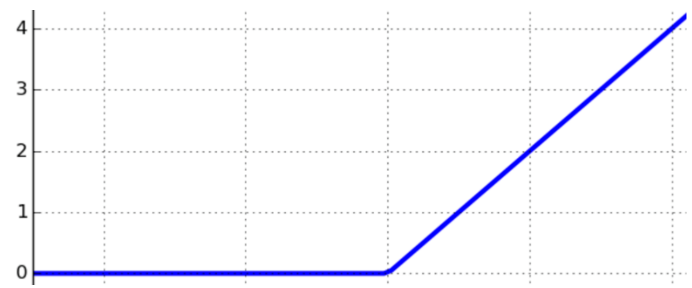
https://medium.com/@curiousily/tensorflow-for-hackers-part-iv-neural-network-from-scratch-1a4f504dfa8
https://hackernoon.com/visualizing-parts-of-convolutional-neural-networks-using-keras-and-cats-5cc01b214e59

Deep Neural Network

- The output of each layer is defined by **activation function**
  - Sigmoid, tanh, or ReLU
- Similar to threshold in human neuron



Sigmoid function

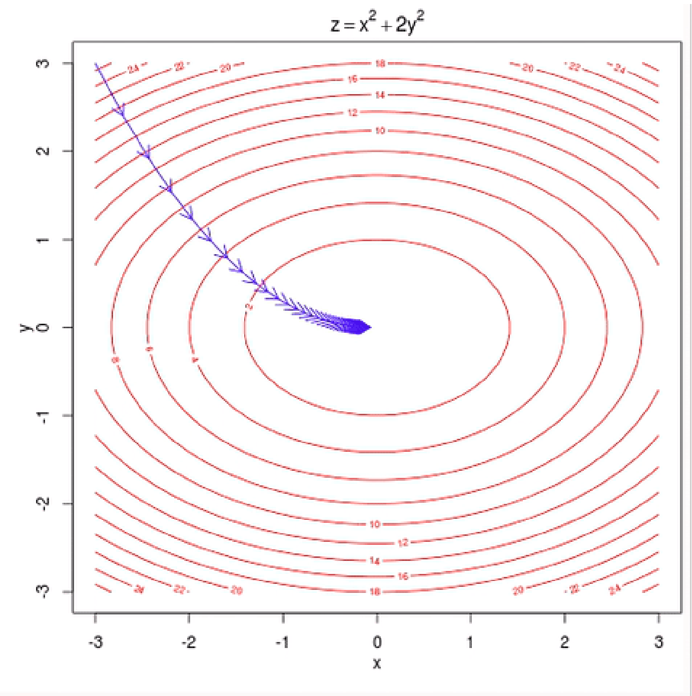


ReLU function

https://en.wikipedia.org/wiki/Sigmoid_function
https://www.oreilly.com/library/view/deep-learning-with/9781787128422/8e232ec5-c2b8-48ee-b4ea-294d1e344533.xhtml

Deep Neural Network

- **Loss function** is used for making DNN to bring correct answer
- Change weight and bias values in each layer for minimizing the cost
  - By **gradient descent** algorithm

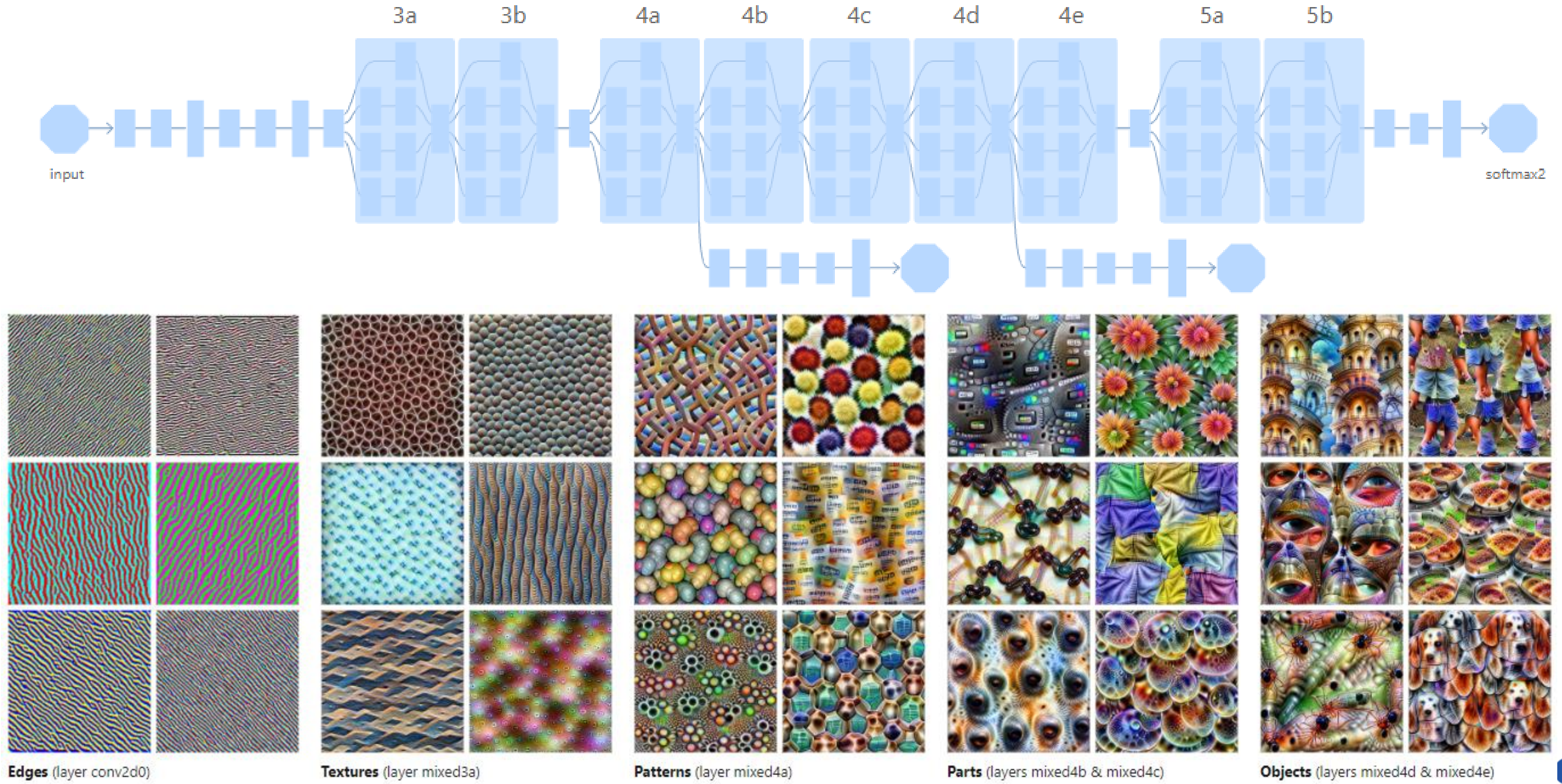


Decreasing cost
By gradient descent

https://giphy.com/gifs/gradient-6QlTwkigqg4yk
Stanford University CS231n

Visualization of neural network

- It understands images from **fine-grained to coarse**



Olah, et al., "Feature Visualization", Distill, 2017.

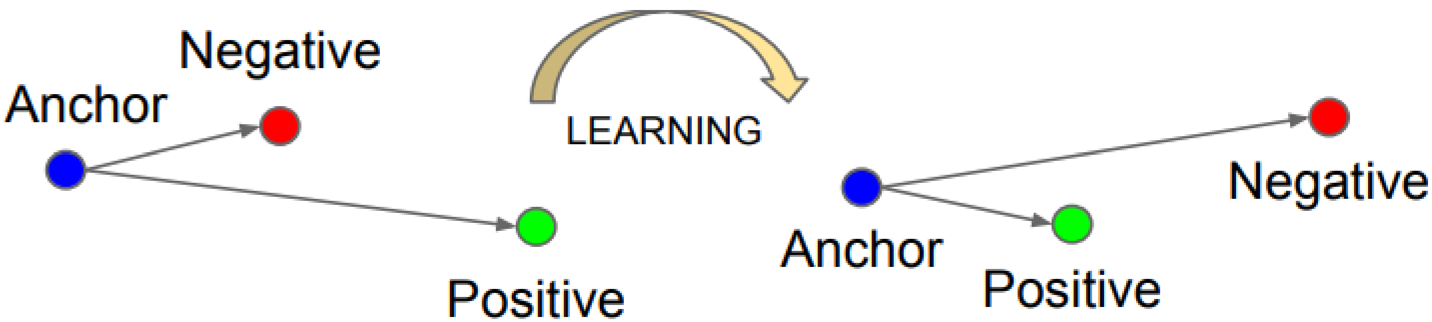
Concept of triplet loss function

- Use **(Anchor, positive, negative)** triplet
- **Minimize** the distance between anchor & positive
- **Maximize** the distance between anchor & negative

 Schroff, et al., "Facenet: A unified embedding for face recognition and clustering", ICCV '15.

Concept of triplet loss function

- Figure and equation of triplet loss function



$$L = \sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

 Schroff, et al., "Facenet: A unified embedding for face recognition and clustering", ICCV '15.

Attention modeling

- Makes neural net to generate **attention mask**
- Can concentrate only on import part

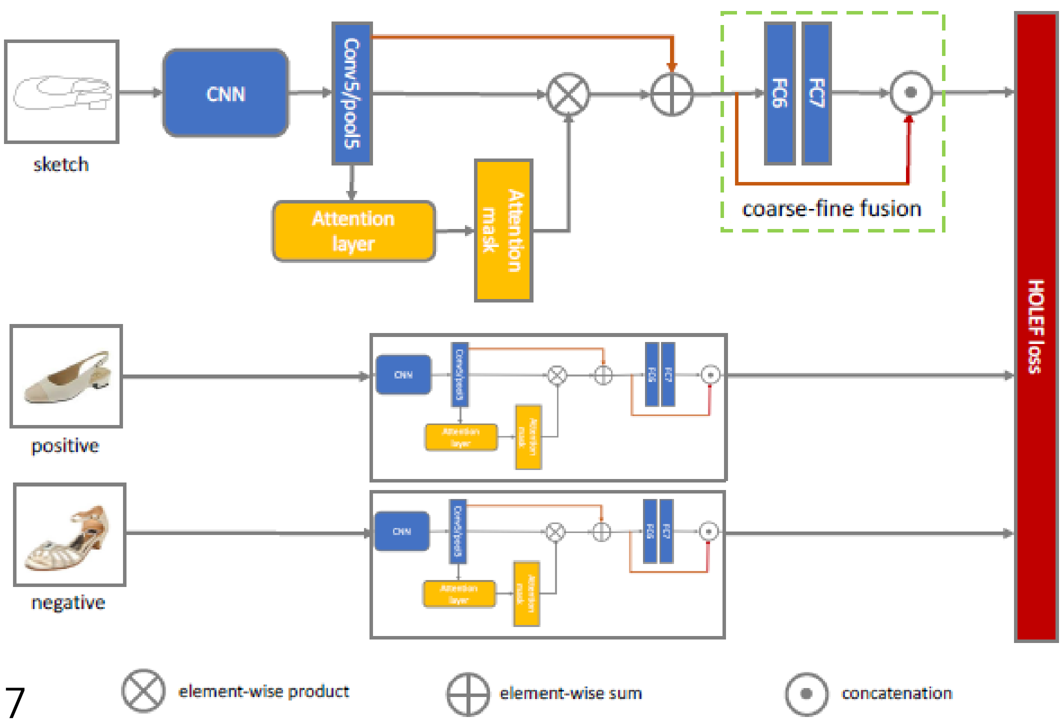
Usage of attention masks in image captioning

Anderson, et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering", CVPR '17.

Main idea

Architecture of proposed model

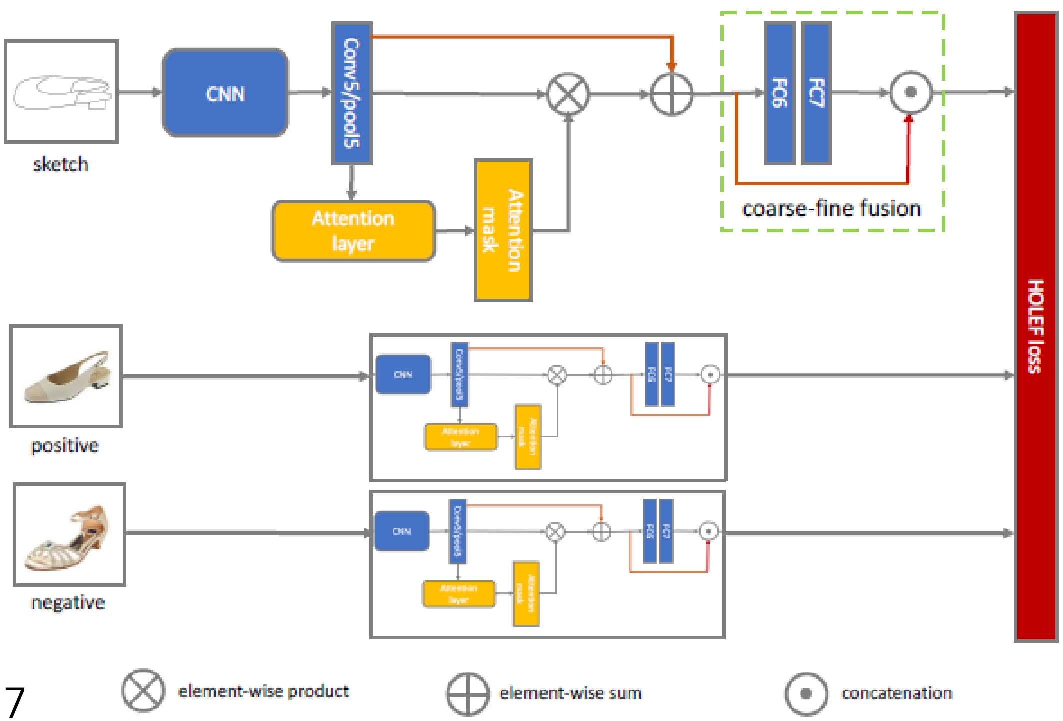
- Sketch, (positive, negative) photos as input
- Attention modeling



Song et al., ICCV '17

Architecture of proposed model

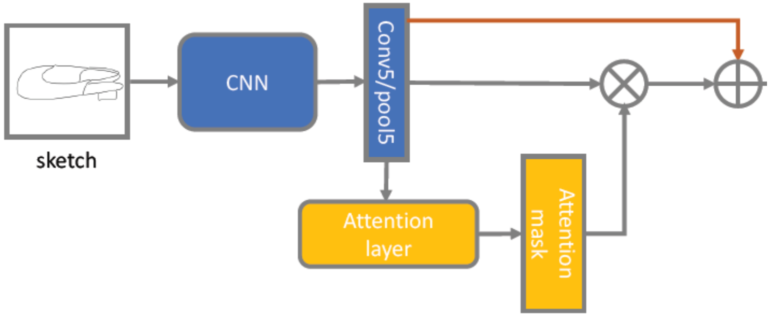
- Coarse-fine fusion
- Triplet loss with a High-order Energy function



Song et al., ICCV '17

Attention modelling

- Use **Conv5 output** for input feature vector
- Calculate **attention score** $\boldsymbol{s_{i,j}}$ from the feature vector
- Generate **attention mask** $\boldsymbol{\alpha_{i,j}}$ using attention score



$$s_{i,j} = F_{att}(f_{i,j}; \boldsymbol{W}_a),$$
$$\alpha_{i,j} = softmax(s_{i,j}),$$

Song et al., ICCV '17

Visualized attention mask

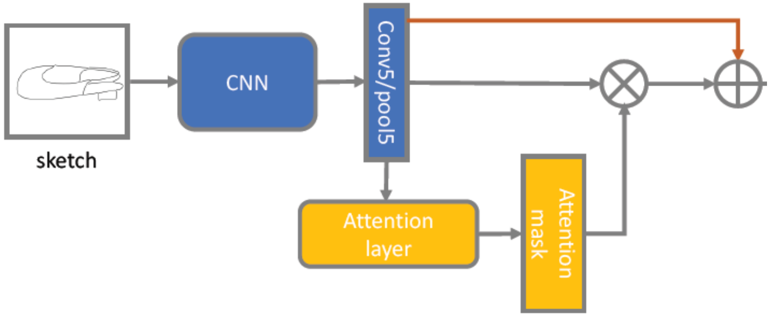
- Showing attention mask for matching sketch & photo

Figure 4. Visualisation of attention masks of sample photo-sketch pairs in all three categories.

Song et al., ICCV '17

Attention modelling

- Element-wise product of $\boldsymbol{\alpha_{i,j}}$ and $\boldsymbol{f_{i,j}}$ for **applying attention mask**
- Element-wise sum of $\boldsymbol{f_{i,j}^{att}}$ and $\boldsymbol{f_{i,j}}$ for **preserving original feature information**

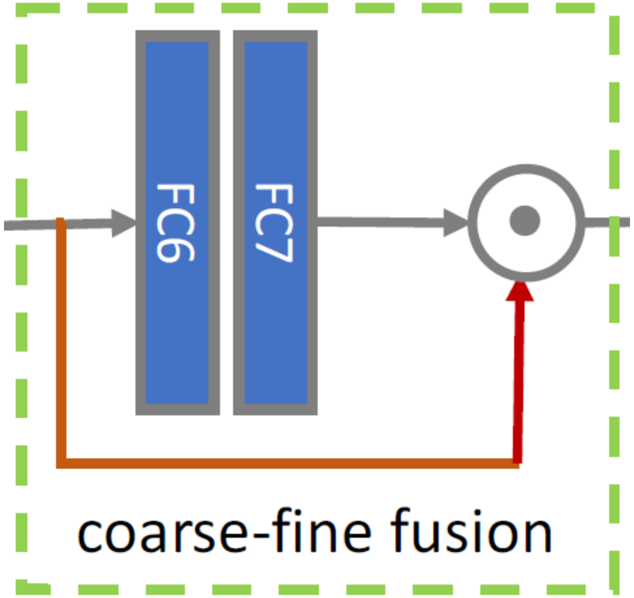


$$s_{i,j} = F_{att}(f_{i,j}; \boldsymbol{W}_a),$$
$$\alpha_{i,j} = softmax(s_{i,j}),$$
$$f_{i,j}^{att} = \alpha_{i,j} \odot f_{i,j}$$
$$\boldsymbol{f}_s^{att} = \boldsymbol{f} + \boldsymbol{\alpha} \odot \boldsymbol{f}$$

Song et al., ICCV '17

Coarse-fine Fusion

- FC7 layer output tends to only have coarse information
- Fuse f_s^{att} and FC7 layer output by concatenation
  - Preserving fine-grained information



Song et al., ICCV '17

KAIST

Triplet Loss w/ Higher-order Energy Function

- Using for overcoming **misalignment** between sketch & photos at Triplet loss calculation
- Compute the **2nd order feature difference** using outer subtraction

$$F_\theta(s) \ominus F_\theta(p) = \begin{bmatrix} {F_\theta}^1(s) \\ {F_\theta}^2(s) \\ {F_\theta}^3(s) \end{bmatrix} \ominus \begin{bmatrix} {F_\theta}^1(p) \\ {F_\theta}^2(p) \\ {F_\theta}^3(p) \end{bmatrix}$$

$$= \begin{bmatrix} {F_\theta}^1(s) - {F_\theta}^1(p) & {F_\theta}^1(s) - {F_\theta}^2(p) & {F_\theta}^1(s) - {F_\theta}^3(p) \\ {F_\theta}^2(s) - {F_\theta}^1(p) & {F_\theta}^2(s) - {F_\theta}^2(p) & {F_\theta}^2(s) - {F_\theta}^3(p) \\ {F_\theta}^3(s) - {F_\theta}^1(p) & {F_\theta}^3(s) - {F_\theta}^2(p) & {F_\theta}^3(s) - {F_\theta}^3(p) \end{bmatrix}$$

Example of outer subtraction

$$\mathcal{D}_H(F_\theta(s), F_\theta(p)) = \sum (F_\theta(s) \ominus F_\theta(p))^{\circ 2} \odot \boldsymbol{W}$$

Proposed energy-function

$$\begin{aligned} L_\theta\left(s, p^+, p^-\right) = \max(0, \Delta + \mathcal{D}_H\left(F_\theta\left(s\right), F_\theta\left(p^+\right)\right) \\ - \mathcal{D}_H\left(F_\theta\left(s\right), F_\theta\left(p^-\right)\right)) + \lambda \|\boldsymbol{W} - \boldsymbol{I}\|_1 \\ + \lambda \|\boldsymbol{W} - \boldsymbol{I}\|_F, \end{aligned}$$

Proposed triplet loss

Song et al., ICCV '17

Newly made dataset – Handbag

- Contains 568 sketch-photo pairs

Figure 3. Examples of newly collected Handbag dataset.

Song et al., ICCV '17

Experiment & Results

Experiment setup

- Experiment on 3 dataset
  - **QMUL-Shoe, QMUL-Chair, and Handbag**
- Classical methods and the first end-to-end deep model for SBIR as baselines
- Ablation studies for figuring out the contribution of each components

Song et al., ICCV '17

Results

- Comparative results against baselines
  - Acc.@1 and Acc.@10

QMUL-Shoe	acc.@1	acc.@10
HOG-BoW + rankSVM	17.39%	67.83%
Dense-HOG + rankSVM	24.35%	65.22%
ISN Deep + rankSVM	20.00%	62.61%
Triplet SN [46]*	52.17 %	92.17 %
Our model	**61.74**%	**94.78**%

QMUL-Chair	acc.@1	acc.@10
HOG-BoW + rankSVM	28.87%	67.01%
Dense-HOG + rankSVM	52.57%	93.81%
ISN Deep + rankSVM	47.42%	82.47%
Triplet SN [46]*	72.16 %	**98.96** %
Our model	**81.44**%	95.88%

Our Handbag	acc.@1	acc.@10
HOG-BoW + rankSVM	2.38%	10.71%
Dense-HOG + rankSVM	15.47%	40.48%
ISN Deep + rankSVM	9.52%	44.05%
Triplet SN [46]*	39.88%	82.14%
Our model	**49.40**%	**82.74**%

Table 1. Comparative results against baselines. '*' The results of Triplet SN [46] are the updated ones from their project webpage which are higher than the published results due to parameter retuning. The other baseline results are copied from [46] except those on Handbag, which are based on our own implementation.

Song et al., ICCV '17

Results

- Contributions of the different components
  - Base: Triplet SN

QMUL-Shoe	acc.@1	acc.@10
Base	52.17%	92.17%
Base + CFF	58.26%	93.04%
Base + HOLEF	56.52%	88.70%
Full: Base + CFF + HOLEF	**61.74**%	**94.78**%
QMUL-Chair	**acc.@1**	**acc.@10**
Base	72.16%	**98.96**%
Base + CFF	79.38%	95.88%
Base + HOLEF	74.23%	97.94%
Full: Base + CFF + HOLEF	**81.44**%	95.88%
Our Handbag	**acc.@1**	**acc.@10**
Base	39.88%	82.14%
Base + CFF	48.21%	83.33%
Base + HOLEF	40.48%	**83.93**%
Full: Base + CFF + HOLEF	**49.40**%	82.74%

Table 2. Contributions of the different components.

Song et al., ICCV '17

KAIST

Results

- Effectiveness of the attention module
  - Base: Triplet SN

QMUL-Shoe	with attention	without attention
Base	54.78%	52.17%
Base + CFF	58.26%	57.39%
Base + HOLEF	57.39%	56.52%
Our model	61.74%	58.26%

QMUL-Chair	with attention	without attention
Base	74.23%	72.16%
Base + CFF	79.38%	75.25%
Base + HOLEF	75.26%	74.23%
Our model	81.44%	77.32%

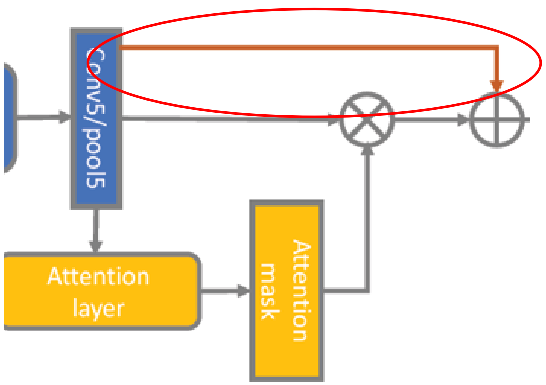
Our Handbag	with attention	without attention
Base	41.07%	39.88%
Base + CFF	48.21%	47.02%
Base + HOLEF	40.48%	40.48%
Our model	49.40%	48.21%

Table 3. Effectiveness of the attention module (acc.@1).

Song et al., ICCV '17

Results

- Effect of shortcut connection in attention module
  - Base: Triplet SN



QMUL-Shoe	with shortcut	without shortcut
Base + attention	54.78%	15.65%
Base + CFF	58.26%	26.96%
Our model	61.74%	27.83%

QMUL-Chair	with shortcut	without shortcut
Base + attention	74.23%	39.18%
Base + CFF	79.38%	48.45%
Our model	81.44%	49.48%

Our Handbag	with shortcut	without shortcut
Base + attention	41.07%	17.26%
Base + CFF	48.21%	24.40%
Our model	49.40%	23.81%

Table 4. Effect of shortcut connection in attention module (acc.@1).

Song et al., ICCV '17

Results

- Comparison on different loss

QMUL-Shoe	acc.@1	acc.@10
Triplet loss with Euclidean	58.26%	93.04%
Triplet loss with Weighted Euclidean	58.26%	93.04%
Triplet loss with Mahalanobis	52.17%	89.57%
Our HOLEF	**61.74%**	**94.78%**

QMUL-Chair	acc.@1	acc.@10
Triplet loss with Euclidean	79.38%	95.88%
Triplet loss with Weighted Euclidean	79.38%	95.88%
Triplet loss with Mahalanobis	78.35%	95.88%
Our HOLEF	**81.44%**	**95.88%**

Our Handbag	acc.@1	acc.@10
Triplet loss with Euclidean	48.21%	**83.33%**
Triplet loss with Weighted Euclidean	48.81%	82.14%
Triplet loss with Mahalanobis	44.64%	79.76%
Our HOLEF	**49.40%**	82.74%

Table 5. Comparison on different losses.

Song et al., ICCV '17

Results

- Retrieval results of baseline and proposed approach
  - Top – proposed, bottom – baseline

Figure 5. Comparison of the retrieval results of our model and Triplet SN [46]. For each example, the top row is our retrieval result with attention mask superimposed on the query sketch, and the bottom row is retrieval result of the same sketch using Triplet SN.

Song et al., ICCV '17

Conclusion

- Main contribution
  - Attention modeling
  - Coarse-Fine fusion
  - New higher-order learnable energy function for triplet loss
- Each component's contribution is proved by ablation study
- Better retrieval result than baseline

Song et al., ICCV '17

KAIST

Thank you!

Quiz #1

- Why the authors use higher-order energy function?
  - A. For faster calculation speed
  - B. To keep both coarse and fine information of images
  - C. For overcoming the misalignment between sketches and photos

Quiz #2

- How the authors overcome the noisy attention mask problem caused by misalignment between sketches and photos?
  - A. Triplet ranking loss
  - B. Fuse attended feature map and original feature map
  - C. Fuse FC7 feature map and attended feature map