Composing Text and Image for Image Retrieval - An Empirical Odyssey

Vo. et al, CVPR 2019

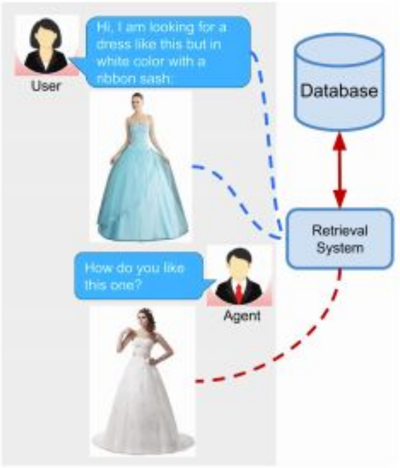
Presented by Mincheul Kim

Table of List

Motivation & Background

Image Retrieval

Task: Image(Input query) + text(describes desired modifications) to the input image



Text query:

Image query:

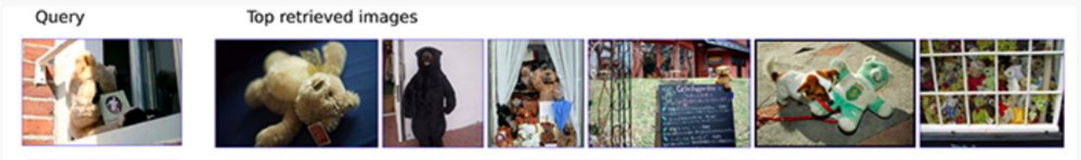
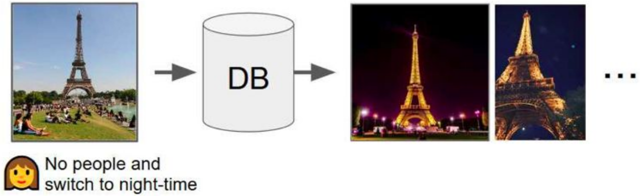


Image + text

Composition query



picture from Naver labs, Europe

Problem

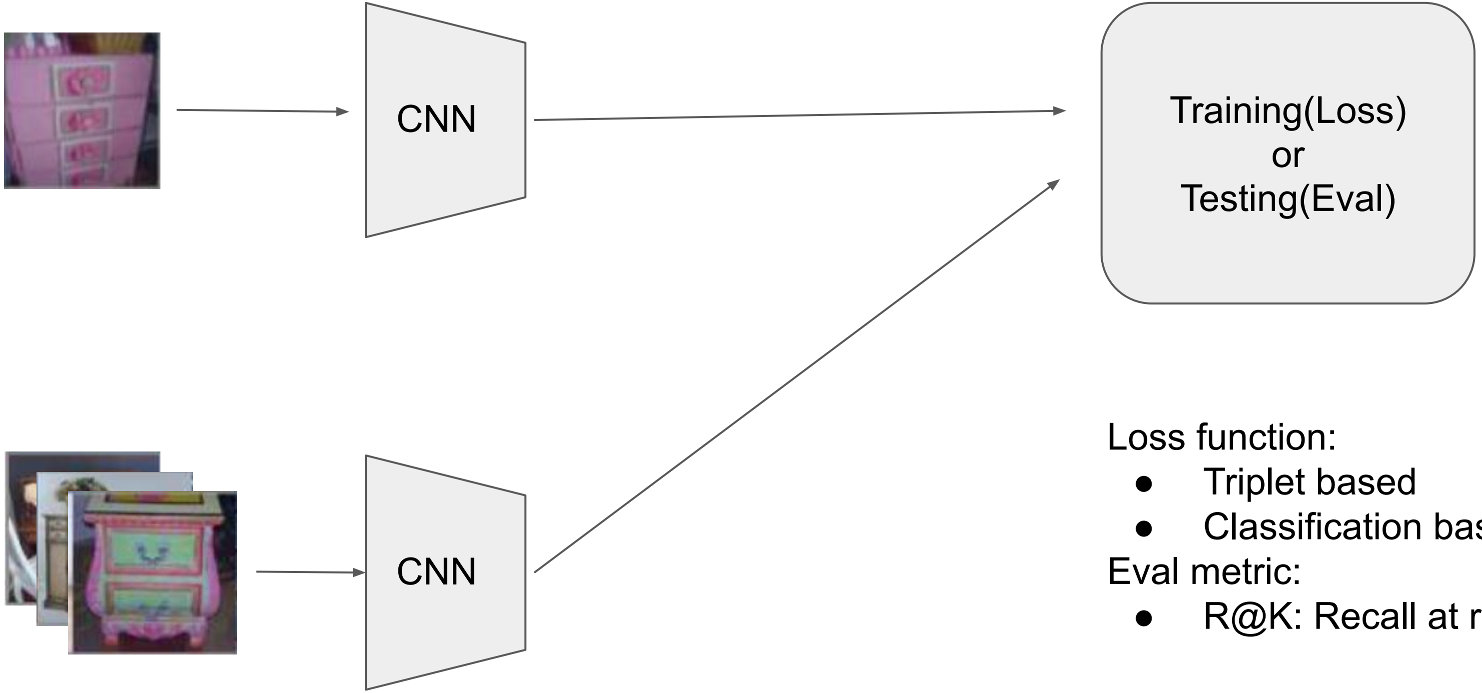
How to get similarity between query and target image?

- Triplet loss, Euclidean, ...

Then, how to represent query with two different modalities?

- Image + text

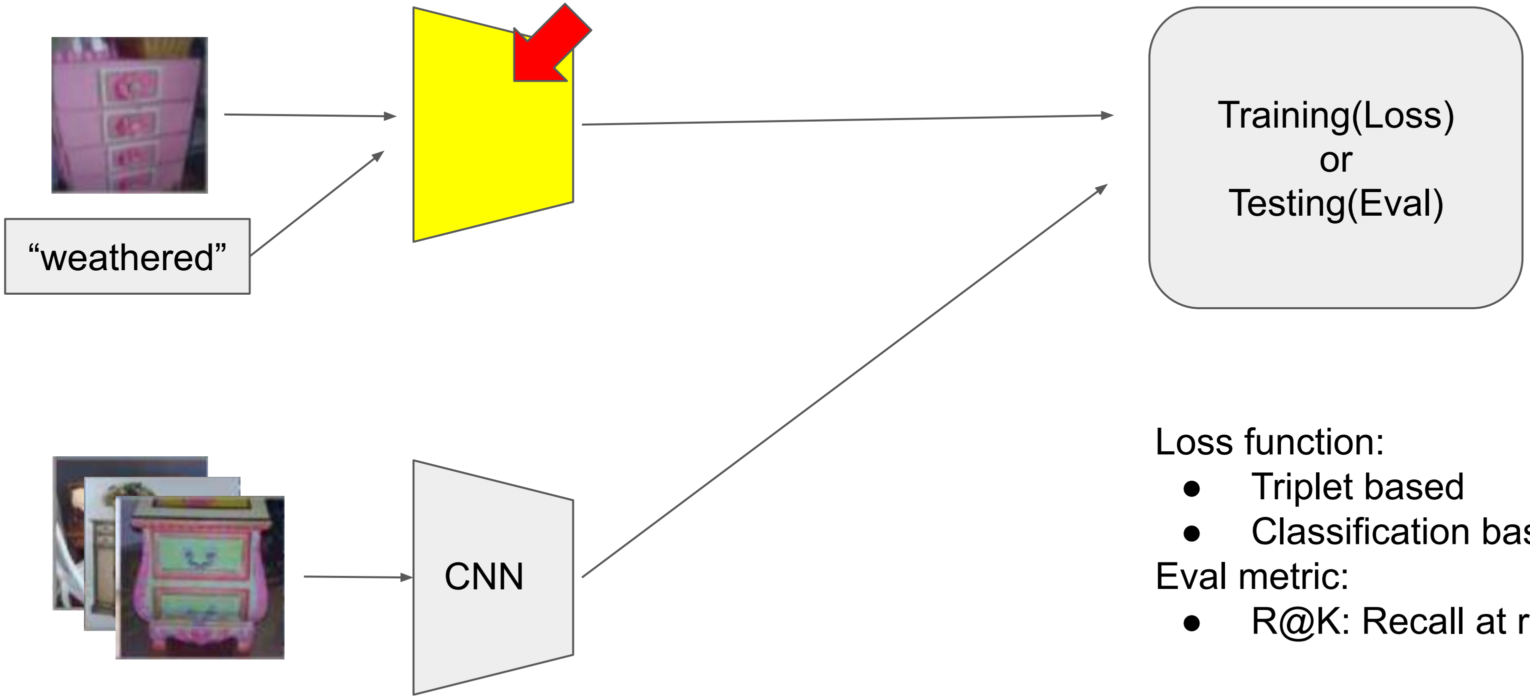
Deep metric Learning



Loss function:
- Triplet based
- Classification based

Eval metric:
- R@K: Recall at rank k

Deep metric Learning



Loss function:

- Triplet based
- Classification based

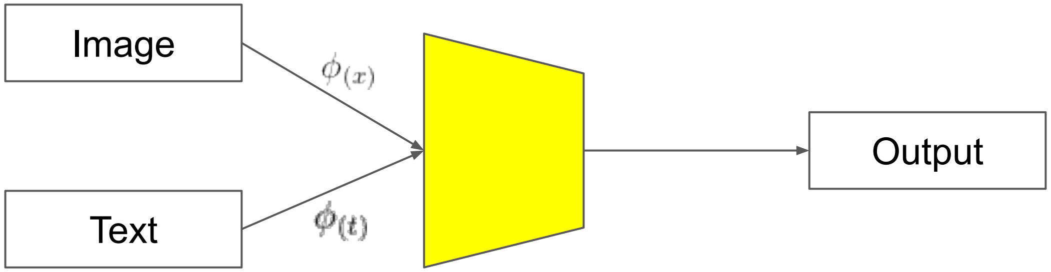
Eval metric:

- R@K: Recall at rank k

Composition of Image and Text

Baseline:

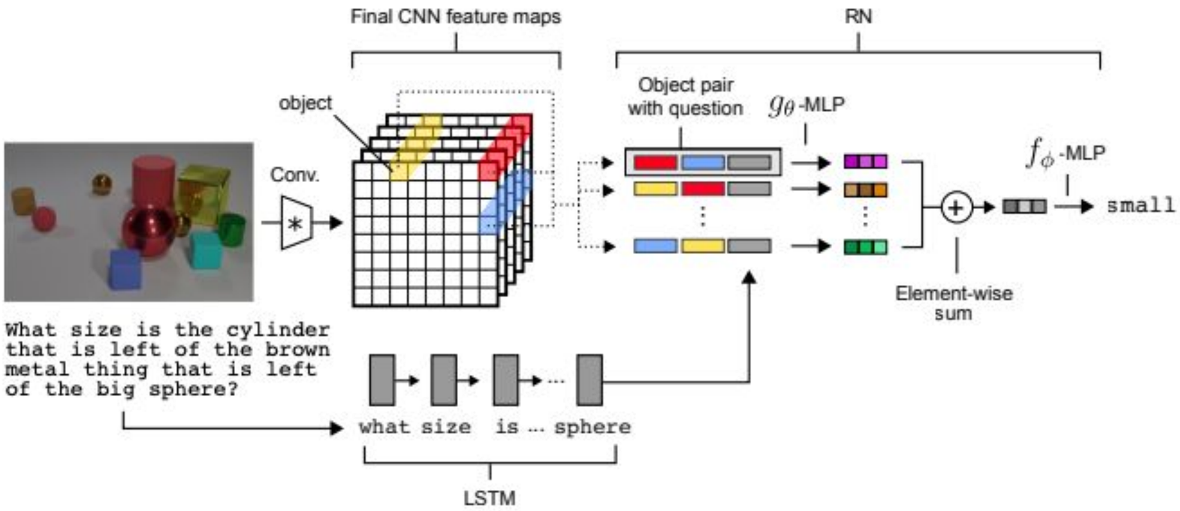
- Encode image and text separately, then perform feature fusion
  - Concatenate (+ feed forward network)
- Captioning and VQA architectures
  - Show n Tell, Relationship Model, FILM



Composition of Image and Text (VQA)

Relationship : concatenate image(CNN) and text(LSTM)

MLP to learn the cross-modal relationships



A. Santoro et al,. A simple neural network module for relational reasoning. In NIPS, 2017

Composition of Image and Text (VQA)

FiLM : text(RNN) cascaded after image(CNN)

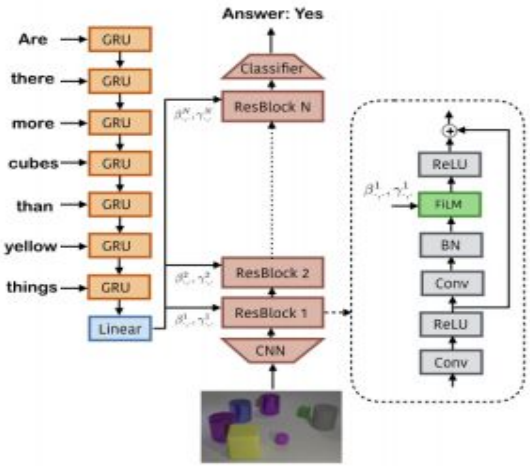


Figure 3: The FiLM generator (left), FiLM-ed network (middle), and residual block architecture (right) of our model.

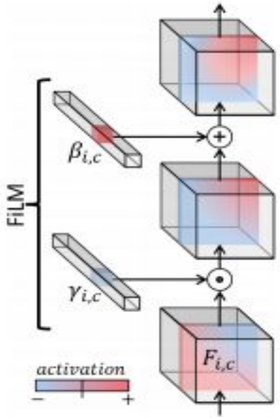


Figure 2: A single FiLM layer for a CNN. The dot signifies a Hadamard product. Various combinations of γ and β can modulate individual feature maps in a variety of ways.

E. Perez et al,. Film: Visual reasoning with a general conditioning layer. 2018

Composition of Image and Text (VQA)

Show and Tell : Train an LSTM to encode both image and text

- O. Vinyals et al,. Show and tell: A neural image caption generator. In CVPR, 2015.

Parameter Hashing : text feature is hashed into transformation matrix

replace weights of FC layers of image(CNN)

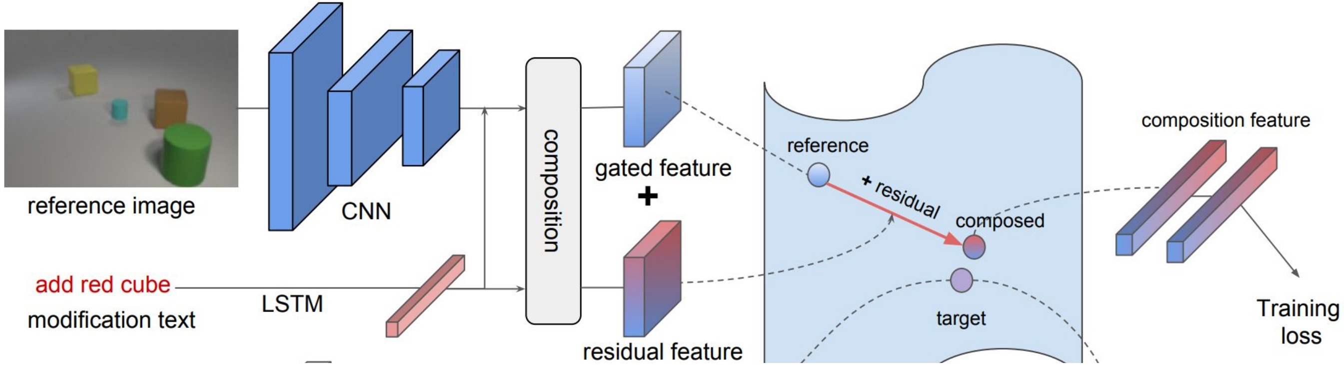
- H. Noh et al,. Image question answering using convolutional neural network with dynamic parameter prediction. In CVPR, 2016

Method

TIRG(Text Image Residual Gating)

Image and text composition mechanism:

- Encode image and text features
- Instead of creating a brand new output (like feature fusion), “modify” the input image feature and return it
- resulting feature still “live in” the same space as target image



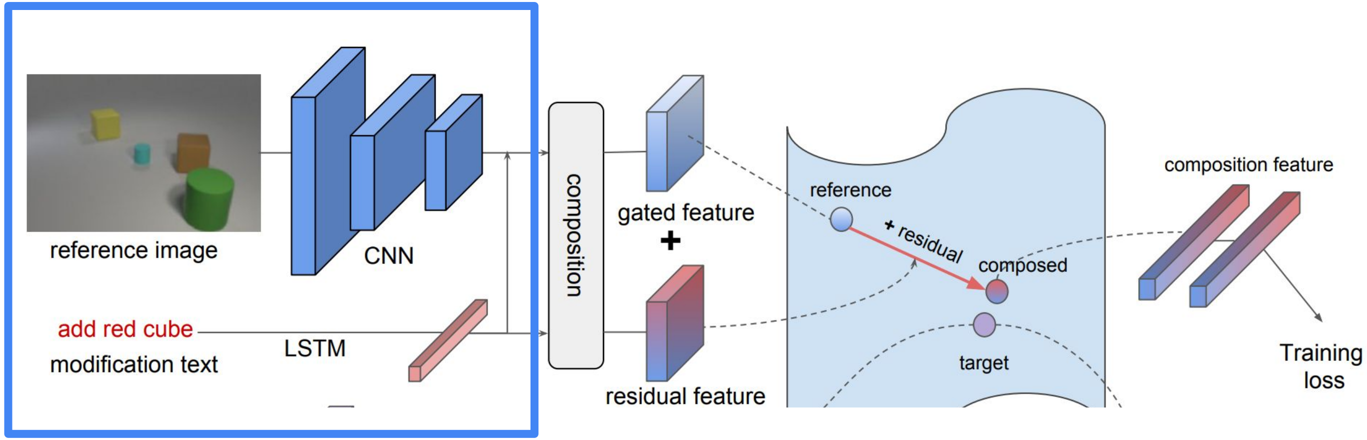
TIRG(Text Image Residual Gating)

Encoding features:

- Reference image: ResNet-17 CNN
- Modification text: LSTM

$\phi_x \in \mathbb{R}^{W \times H \times C}$

$\phi_t \in \mathbb{R}^d$

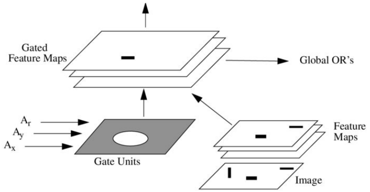


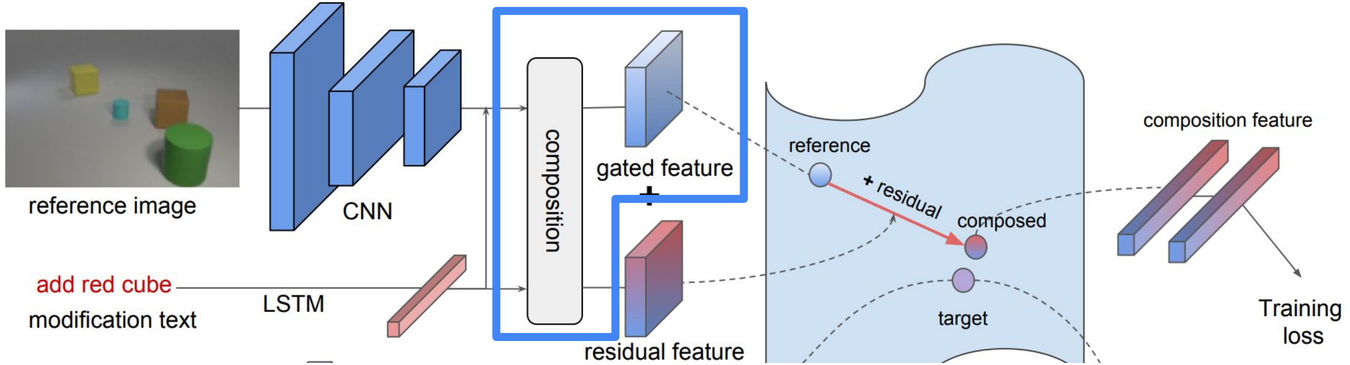
TIRG(Text Image Residual Gating)

Gating connection:

- Establish input image feature as reference to output composition feature
- Network to control what visual information should be enhanced according to the text

$$f_{\text{gate}}(\phi_x, \phi_t) = \sigma(W_{g2} * \text{RELU}(W_{g1} * [\phi_x, \phi_t]) \odot \phi_x$$



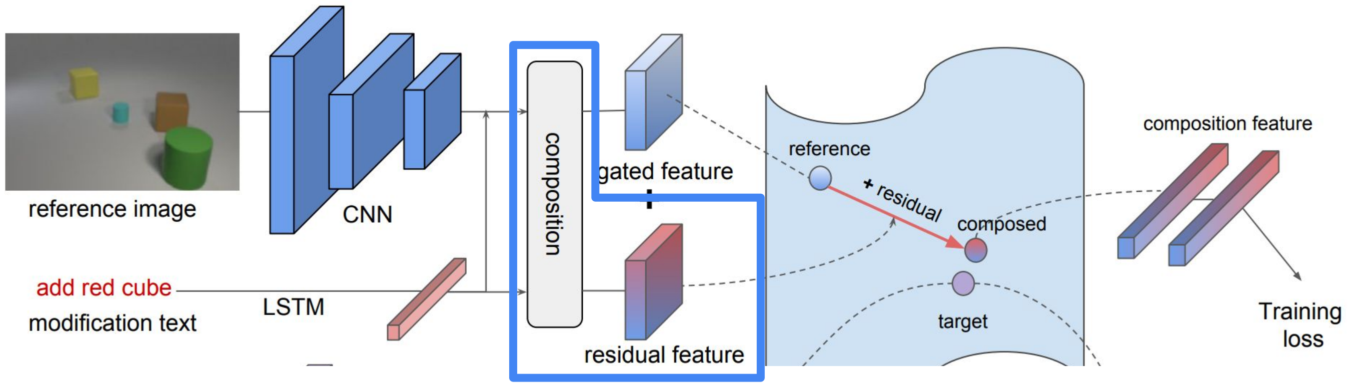


TIRG(Text Image Residual Gating)

Residual connection:

- represents the modification or “walk” in this feature space
- Learns similarity between gated features and target image features

$$f_{\text{res}}(\phi_x, \phi_t) = W_{r2} * \text{RELU}(W_{r1} * ([\phi_x, \phi_t]))$$

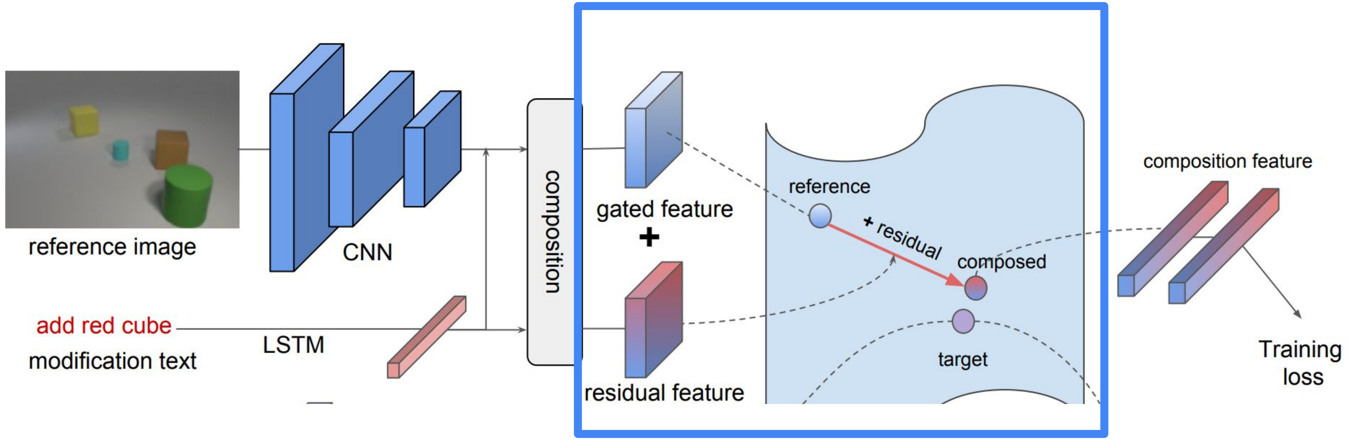


TIRG(Text Image Residual Gating)

Feature composition:

- Combine two features
- Start as working image retrieval, then gradually learn meaningful modification

$$\phi_{xt}^{rg} = w_g f_{\text{gate}}(\phi_x, \phi_t) + w_r f_{\text{res}}(\phi_x, \phi_t)$$



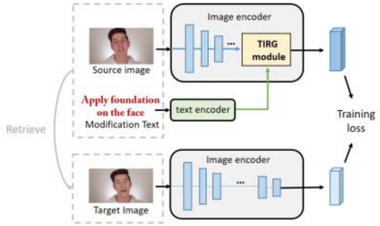
Similarity Measure(Training)

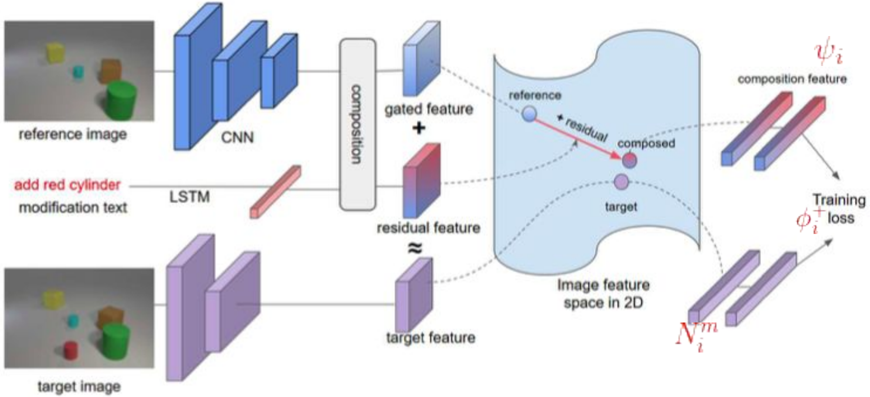
Objective: push closer features of the “modified” and target image

Batch Classification Loss:

$$L = \frac{-1}{MB} \sum_{i=1}^{B} \sum_{m=1}^{M} \log\{ \frac{\exp\{\kappa(\psi_i, \phi_i^+)\}}{\sum_{\phi_j \in \mathcal{N}_i^m} \exp\{\kappa(\psi_i, \phi_j)\}} \}$$

- B: training minibatch
- M: iteration (B/K)
- ψ_i : final representation of image-text query
- ϕ_i^+ : target image(positive feature)
- N_i^m: possible set(ϕ_i^+ +K-1 negative e.g)





Experiments

Experiment configuration

Datasets: Fashin200k, MIT-States, CSS

Metric: R@K (recall at rank k)

Image encoder: ResNet-17 pretrained on ImageNet (output feature size = 512)

Text encoder: LSTM of random initial weight (hidden size = 512)

Training is run for 150k iteration with a start learning rate 0.01

Fashion200k

~200k images of fashion products

Category labels : dress, top, pants, skirt, jacket

Compact attribute-like product description
  e.g. black jacket

Modification text: one different word

Method	R@1	R@10	R@50
Han *et al.* [12]	6.3	19.9	38.3
Image only	3.5	22.7	43.7
Text only	1.0	12.3	21.8
Concatenation	$11.9^{\pm 1.0}$	$39.7^{\pm 1.0}$	<u>62.6</u>$^{\pm 0.7}$
Show and Tell	$12.3^{\pm 1.1}$	$40.2^{\pm 1.7}$	$61.8^{\pm 0.9}$
Param Hashing	$12.2^{\pm 1.1}$	$40.0^{\pm 1.1}$	$61.7^{\pm 0.8}$
Relationship	$13.0^{\pm 0.6}$	<u>40.5</u>$^{\pm 0.7}$	$62.4^{\pm 0.6}$
MRN	<u>13.4</u>$^{\pm 0.4}$	$40.0^{\pm 0.8}$	$61.9^{\pm 0.6}$
FiLM	$12.9^{\pm 0.7}$	$39.5^{\pm 2.1}$	$61.9^{\pm 1.9}$
TIRG	**14.1**$^{\pm 0.6}$	**42.5**$^{\pm 0.7}$	**63.8**$^{\pm 0.8}$

Table 1. Retrieval performance on Fashion200k. The best number is in bold and the second best is underlined.

MIT-States

~60k images

245 nouns and 115 adjectives

Object/noun label + state/adjective label
  e.g. frozen cheese, new table clock

Modification text: state

Method	R@1	R@5	R@10
Image only	$3.3^{\pm 0.1}$	$12.8^{\pm 0.2}$	$20.9^{\pm 0.1}$
Text only	$7.4^{\pm 0.4}$	$21.5^{\pm 0.9}$	$32.7^{\pm 0.8}$
Concatenation	$11.8^{\pm 0.2}$	$30.8^{\pm 0.2}$	$42.1^{\pm 0.3}$
Show and Tell	$11.9^{\pm 0.1}$	$31.0^{\pm 0.5}$	$42.0^{\pm 0.8}$
Att. as Operator	$8.8^{\pm 0.1}$	$27.3^{\pm 0.3}$	$39.1^{\pm 0.3}$
Relationship	$\mathbf{12.3}^{\pm 0.5}$	$\mathbf{31.9}^{\pm 0.7}$	$\underline{42.9}^{\pm 0.9}$
MRN	$11.9^{\pm 0.6}$	$30.5^{\pm 0.3}$	$41.0^{\pm 0.2}$
FiLM	$10.1^{\pm 0.3}$	$27.7^{\pm 0.7}$	$38.3^{\pm 0.7}$
TIRG	$\underline{12.2}^{\pm 0.4}$	$\mathbf{31.9}^{\pm 0.3}$	$\mathbf{43.1}^{\pm 0.3}$

Table 2. Retrieval performance on MIT-States.

CSS

~34k images

Modification text: add/remove/change + color, shape, size

e.g. add red sphere to top-left

Two retrieval setting: 3D & 2D query image

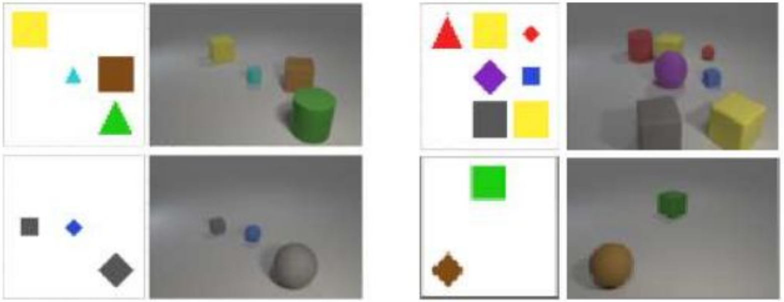


Figure 5. Example images in our CSS dataset. The same scene are rendered in 2D and 3D images.

Method	3D-to-3D	2D-to-3D
Image only	6.3	6.3
Text only	0.1	0.1
Concatenate	$60.6^{\pm 0.8}$	27.3
Show and Tell	$33.0^{\pm 3.2}$	6.0
Parameter hashing	$60.5^{\pm 1.9}$	31.4
Relationship	$62.1^{\pm 1.2}$	30.6
MRN	$60.1^{\pm 2.7}$	26.8
FiLM	$65.6^{\pm 0.5}$	43.7
TIRG	**73.7**$^{\pm 1.0}$	**46.6**

Table 4. Retrieval performance (R@1) on the CSS Dataset using 2D and 3D images as the query.

CSS

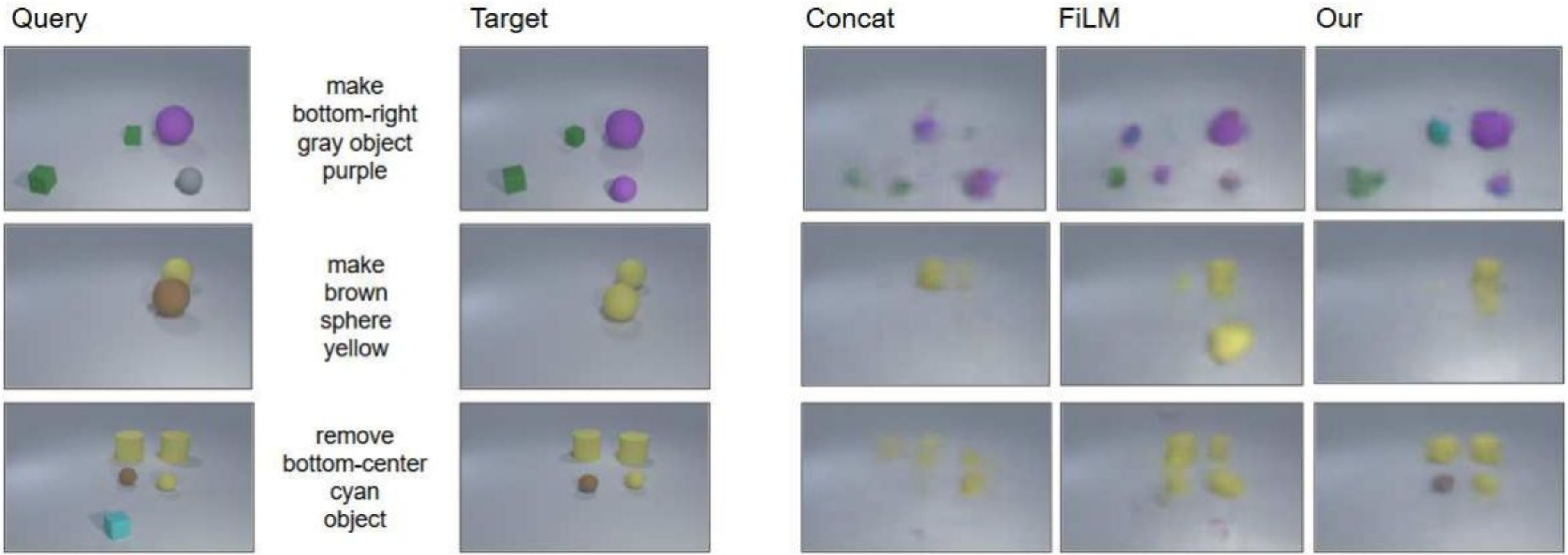


Figure 7. Reconstruction images from the learned composition features.

Summary

Contribution

Study feature composition for image retrieval, and proposed a new method

- “modifying” reference image feature with gating & residual connection

Create a new data set, CSS

- enables controlled experiments of image retrieval using text and image queries

Limitation

- Limitation of text manipulation
  - text descriptions are more subjective than using absolute attribute values which can sometimes be problematic
  - using a text description to define an image may not always result in the desired image as the same text can correspond to multiple images
- Direct combines text feature of the entire sentence with image feature
  - Requires detailed understanding of linguistic information of the word in different region
- Many parts that need explanation are missing
  - Why LSTM is used for text encoding? other like RNN-based, BERT?
  - Missing enough explanation in method (e.g. gating, residual, …)
- Lacks of various evaluation metric
  - computation time, memory size, ...