Composed Query Image Retrieval Using Locally Bounded Features

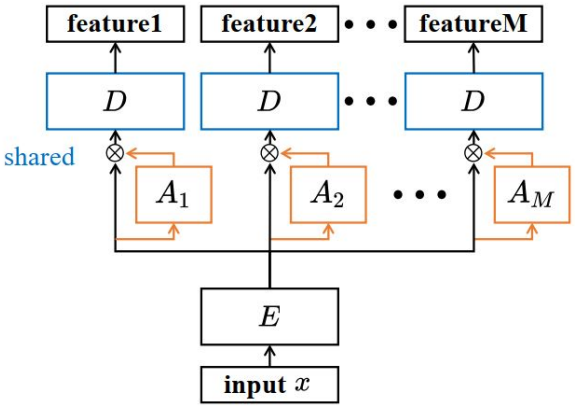
Hosseinzadeh and Wang., CVPR 2020
Presented by Mincheul Kim

Table of List

Review: Attention-based Ensemble for Deep Metric Learning

Use multiple models to obtain better performance

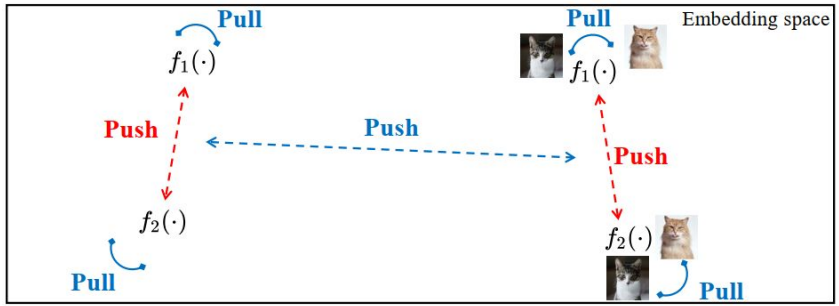
- Apply Attention for high-performance
- Propose Divergence Loss for model diversity



< Our Architecture >

Total Loss: Pairwise Loss + Divergence Loss

$$L = \sum_{m=1}^{M} L_{pair,(m)} + \lambda_{div} L_{div}$$



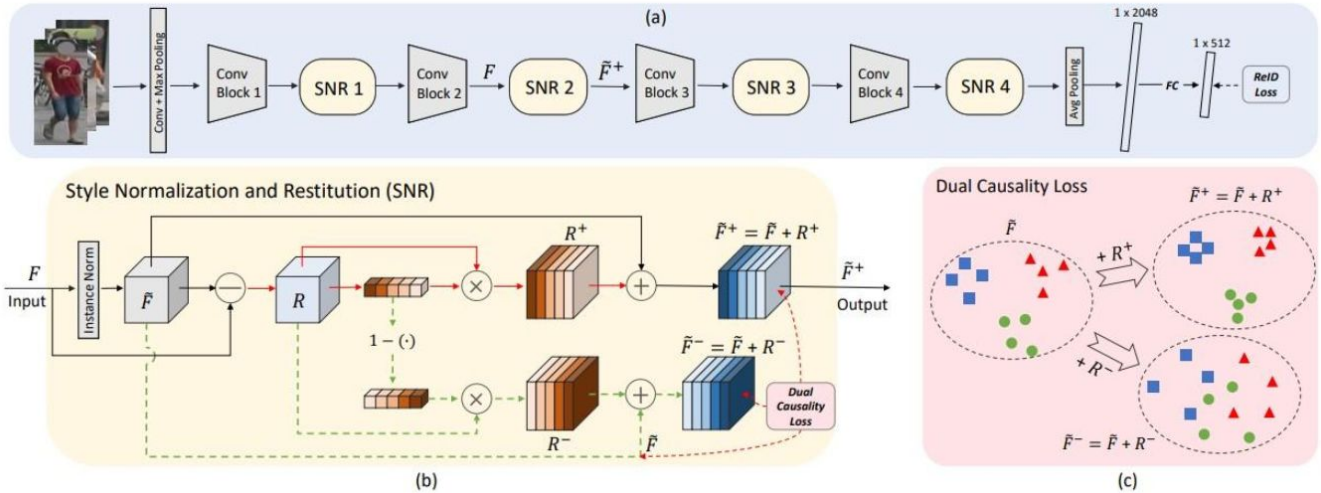
Compositional Learning of Image-Text Query for Image Retrieval, WACV 2021

Review: Style Normalization and Restitution for Generalizable Person Re-identification

Residual feature R : Difference between original and style normalized feature

Restituting it into identity-relevant(+) and identity-irrelevant feature(-)

Dual casualty loss = Clarification loss + destruction loss



Motivation & Background

Background

Composed Query image retrieval

- Query image + text(requested modification)

Previous methods

- Usually consider the image as a whole

Motivation

Modification text usually refers to one or more “entities”

- entities: image that should be changed

Prior works: consider an image as a whole

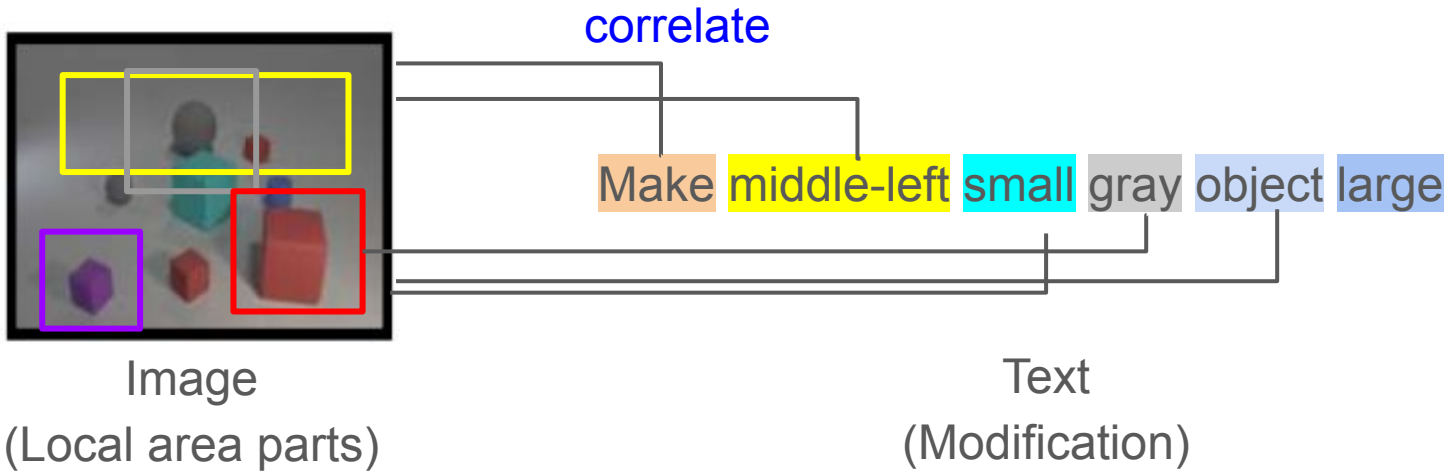
- processing entire image at once using a CNN

Ours: consider the image as a set of local semantic entities

Approach

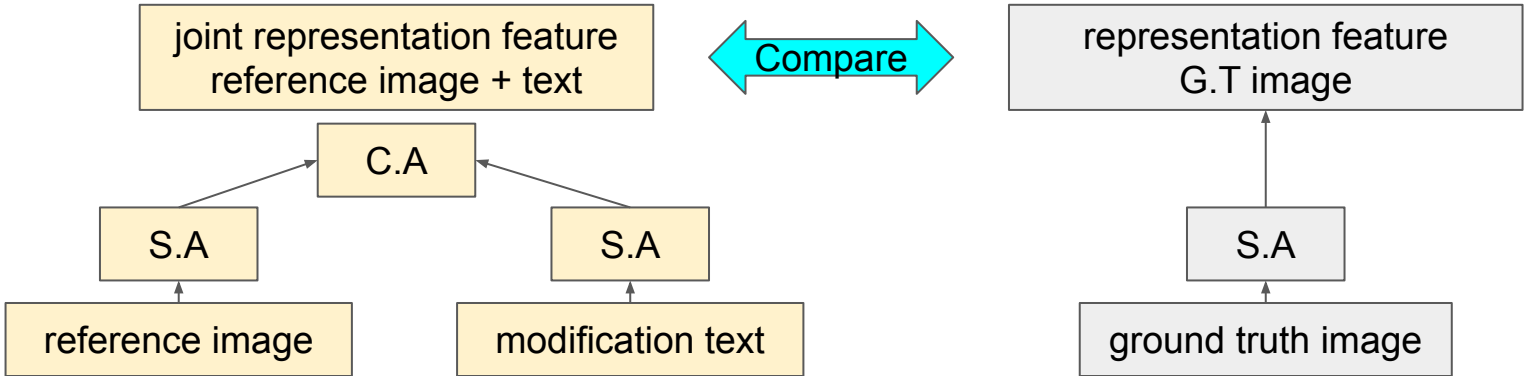
Represents the image using a set of local areas

Explicitly establish relationship between each word(in the modification text) and each area in the image



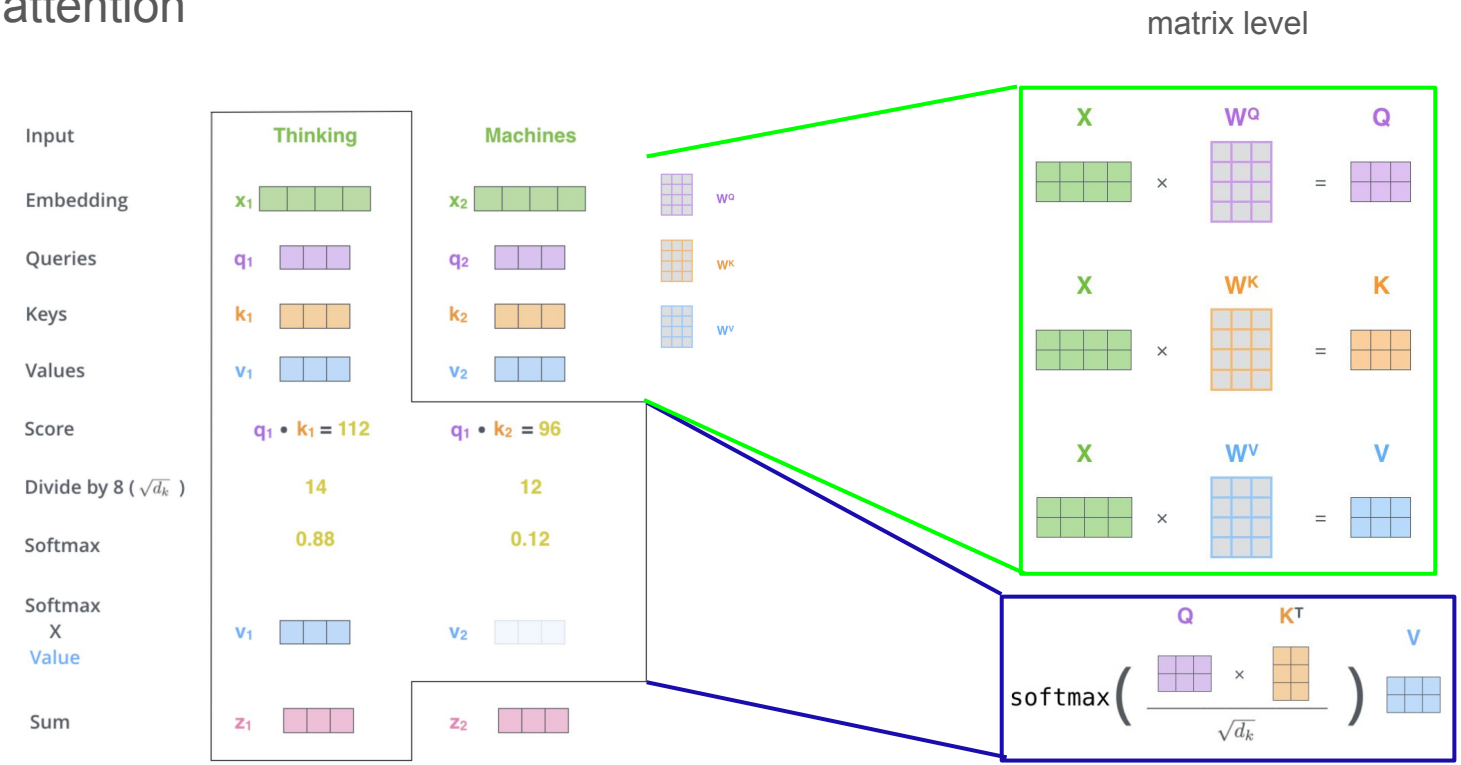
Method

1. Extracts the features for a set of local areas in the images
   - Each of these local regions = “entity”
2. Set of features and modification text processed using separate branches with Self-attention layers
3. Cross-modal module learns a joint representation of the query image and the modification text
   - By leveraging attention mechanism to correlate each word to each entity in the image



Related works

Self attention

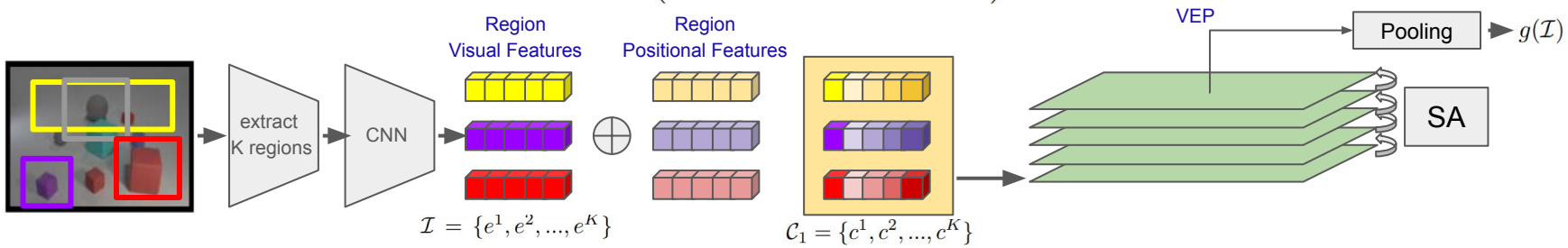


1. Image Representation with Locally Bounded Features

Divide the image into locally bounded entities and process image at region level

1. Region Visual Features
   a. Using Faster R-CNN(pre-trained): extract K regions
   b. Each regions is represented as CNN feature vectors $\mathcal{I} = \{e^1, e^2, ..., e^K\}$
2. Region Positional Features
   a. Composed queries contain positional words(e.g. replace the oval right to circle with …)
   b. Represent layout image(spatial relationships between different objects(region) in the image)
   c. Calculate positional feature vector(normalized information of the i-th region)

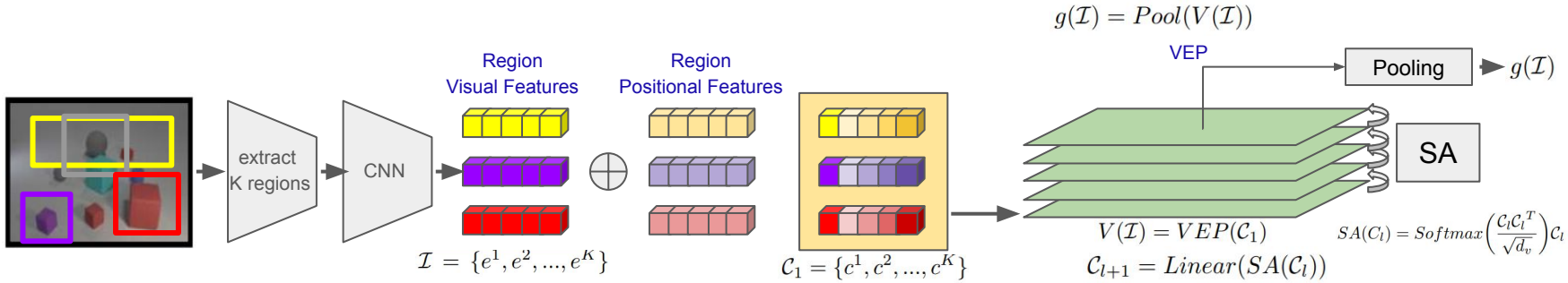
$$p^i = Linear\Big([N(x^i), N(y^i), N(w^i), N(h^i)]\Big)$$



1. Image Representation with Locally Bounded Features

Divide the image into locally bounded entities and process image at region level

3. Image Representation
   a. Average visual and positional features for every region $c^i = Linear(avg(e^i, p^i))$
   b. VEP: Self-attention based multi-layer visual embedding processing module
      i. C1 is the input to first layer of VEP $\mathcal{C}_1 = \{c^1, c^2, ..., c^K\}$
      ii. Get final feature representation of the image
   c. Output of the last layer of VEP is used as image representation $V(\mathcal{I}) \in \mathbb{R}^{K \times d_v}$



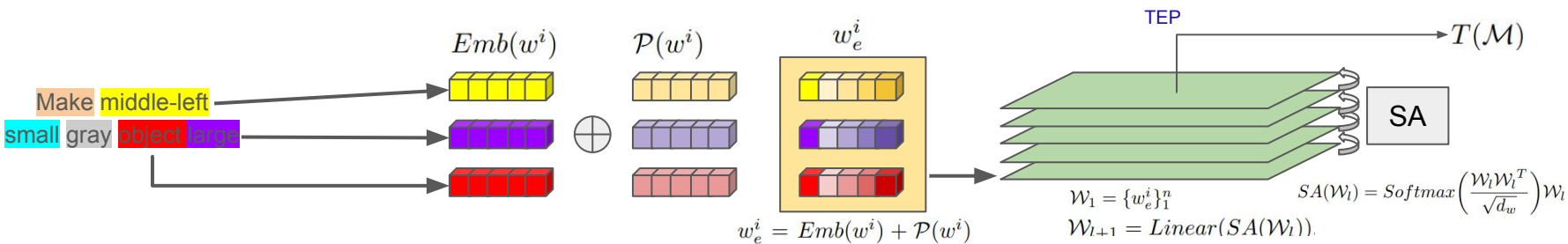
2. Modification Text Features

Process composed query sentence M which is a sequence of n words

- Each word mapped to vector by two seperate embedding layers $Emb(w^i)$, $\mathcal{P}(w^i)$
- Final representation for i-th word in sentence $w_e^i = Emb(w^i) + \mathcal{P}(w^i)$

Initial input to TEP: sequence of word representations $\mathcal{W}_1 = \{w_e^i\}_1^n$

Similar to visual embedding process



3. Feature Fusion

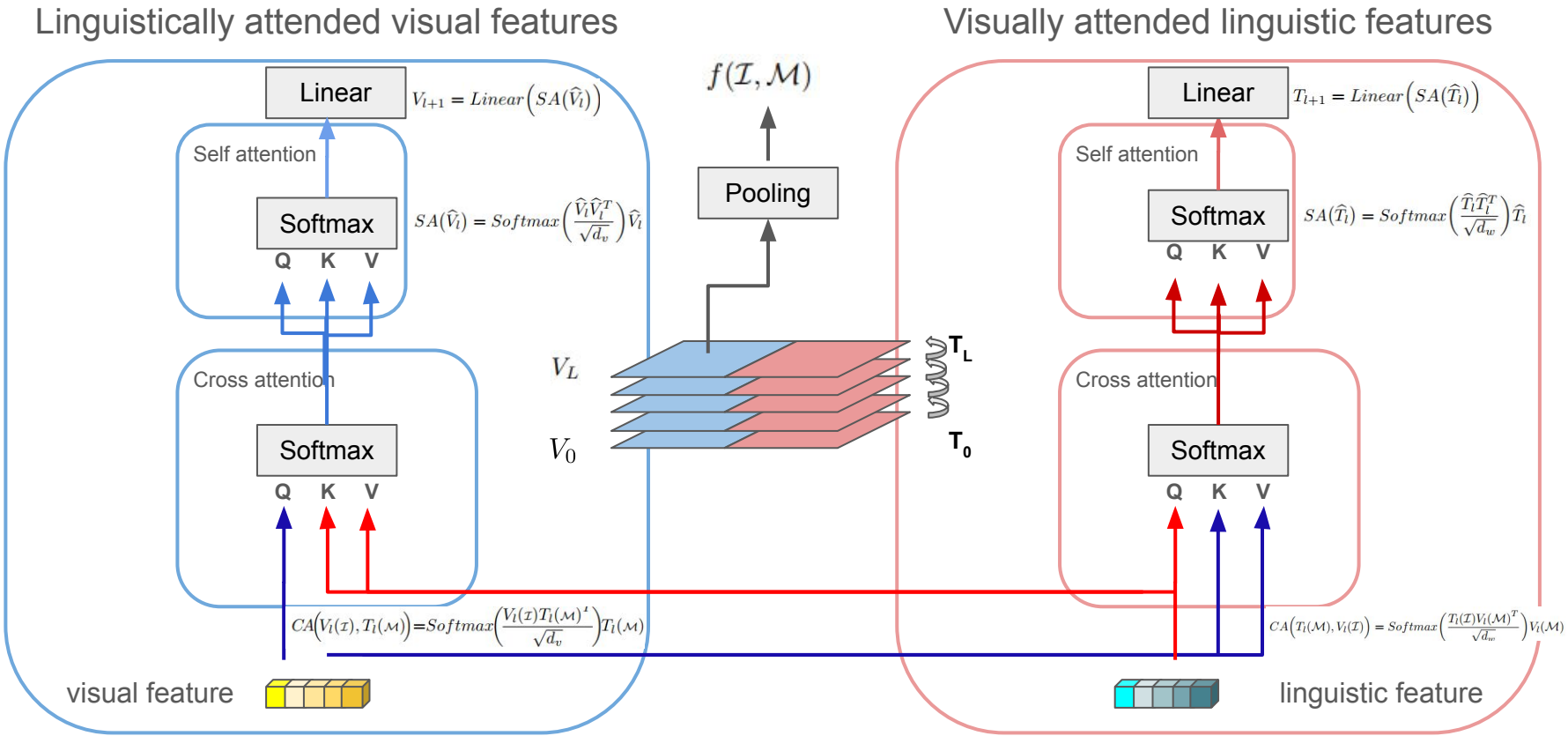
Integrating information from reference image and modification text

- prior(TIRG): Directly combine feature vector of entire query sentence with feature vector of the entire image → Not effective

Intuition: Incorporate cross-modal attention module to fuse there two modalities

- Linguistically attended visual features
- Visually attended language features
- Jointly representation of composed features $f(\mathcal{I}, \mathcal{M})$

3. Feature Fusion



4. Similarity Learning

Task: Learn the model parameters

Loss function $sim(f(\mathcal{I}, \mathcal{M}), g(\mathcal{I}_t)) \gg sim(f(\mathcal{I}, \mathcal{M}), g(\mathcal{I}_{c_i}))$

- Soft triplet loss $\mathcal{L}_{ST} = \sum_{i=1}^{k-1} \log\left(1 + \frac{\exp(sim(f(\mathcal{I}, \mathcal{M}), g(\mathcal{I}_t)))}{\exp(sim(f(\mathcal{I}, \mathcal{M}), g(\mathcal{I}_{c_i})))}\right)$
- Batch classification loss $\mathcal{L}_{BC} = \frac{1}{|B|} \sum_{i=1}^{|B|} -\log\left(\frac{\exp\left(sim(f(\mathcal{I}_i, \mathcal{M}_i), g(\mathcal{I}_{t_i}))\right)}{\sum_{j=1}^{k-1} \exp\left(sim(f(\mathcal{I}_i, \mathcal{M}_i), g(\mathcal{I}_{c_j}))\right)}\right)$

Experiment and Result

Fashion-200k

Outperform other baseline

Better results when K = 36(big) than K = 18(small)

- K = region proposal for each image

Method	*Recall@*		
	K=1	*K=10*	*K=50*
Baselines			
Image only [35]	3.5	22.7	43.7
Text only [35]	1.0	12.3	21.8
Concat [35]	$11.9^{\pm 1.0}$	$39.7^{\pm 1.0}$	$62.6^{\pm 0.7}$
SOTA			
Han et al. [12]	6.3	19.9	38.3
Show and Tell [34]	$12.3^{\pm 1.1}$	$40.2^{\pm 1.7}$	$61.8^{\pm 0.9}$
Param. Hash. [21]	$12.2^{\pm 1.1}$	$40.0^{\pm 1.1}$	$61.7^{\pm 0.8}$
Relationship [26]	$13.0^{\pm 0.6}$	$40.5^{\pm 0.7}$	$62.4^{\pm 0.6}$
FiLM [23]	$12.9^{\pm 0.7}$	$39.5^{\pm 2.1}$	$61.9^{\pm 1.9}$
TIRG [35]	$14.1^{\pm 0.6}$	$42.5^{\pm 0.7}$	$63.8^{\pm 0.8}$
Ours (big)	$\mathbf{17.78}^{\pm 0.5}$	$\mathbf{48.35}^{\pm 0.6}$	$\mathbf{68.5}^{\pm 0.5}$
Ours (small)	$\mathbf{16.26}^{\pm 0.6}$	$\mathbf{46.90}^{\pm 0.3}$	$\mathbf{71.73}^{\pm 0.6}$

Experiment and Result

MIT States & CSS

Method	*Recall@*		
	K=1	*K=5*	*K=10*
Baselines			
Image only [35]	$3.3^{\pm 0.1}$	$12.8^{\pm 0.2}$	$20.9^{\pm 0.1}$
Text only [35]	$7.4^{\pm 0.4}$	$21.5^{\pm 0.9}$	$32.7^{\pm 0.8}$
Concat [35]	$11.8^{\pm 0.2}$	$30.8^{\pm 0.2}$	$42.1^{\pm 0.3}$
SOTA			
Show and Tell [34]	$11.9^{\pm 0.1}$	$31.0^{\pm 0.5}$	$42.0^{\pm 0.8}$
Attribute Op. [20]	$8.8^{\pm 0.1}$	$27.3^{\pm 0.3}$	$39.1^{\pm 0.3}$
Relationship [26]	$12.3^{\pm 0.5}$	$31.9^{\pm 0.7}$	$42.9^{\pm 0.9}$
FiLM [23]	$10.1^{\pm 0.3}$	$27.7^{\pm 0.7}$	$42.9^{\pm 0.9}$
TIRG [35]	$12.2^{\pm 0.4}$	$31.9^{\pm 0.3}$	$41.3^{\pm 0.3}$
Ours (big)	$\mathbf{14.72}^{\pm 0.6}$	$\mathbf{35.30}^{\pm 0.7}$	$\mathbf{46.56}^{\pm 0.5}$
Ours (small)	$\mathbf{14.29}^{\pm 0.6}$	$\mathbf{34.67}^{\pm 0.7}$	$\mathbf{46.06}^{\pm 0.6}$

MIT States

Method	*Recall@*	
	$3D \rightarrow 3D$	$2D \rightarrow 3D$
	K=1	*K=1*
Baselines		
Image only [35]	6.3	6.3
Text only [35]	0.1	0.1
Concat [35]	$60.6^{\pm 0.8}$	27.3
SOTA		
Show & Tell [34]	$33.0^{\pm 3.2}$	6.0
Param Hash. [21]	$60.5^{\pm 1.9}$	31.4
Relation. [26]	$62.1^{\pm 1.2}$	30.6
FiLM [23]	$65.6^{\pm 0.5}$	43.7
TIRG [35]	$73.7^{\pm 1.0}$	46.6
Ours (big)	$\mathbf{79.2}^{\pm 1.2}$	$\mathbf{55.69}^{\pm 0.9}$
Ours (small)	$\mathbf{67.26}^{\pm 1.1}$	$\mathbf{50.31}^{\pm 0.9}$

CSS

Qualitative Examples

MIT States

Change state to sliced

Change state to unripe

Fashion -200k

Replace paisley style to geometric

Replace grey color to pink

CSS

Add grey object

Make middle-left small gray object large

Conclusion

Represent input image as a set of local regions(entities)

Learn a bidirectional correlation between the words in the modification text