浙江大學 Zhejiang University 浙江工商大學 ZHEJIANG GONGSHANG UNIVERSITY Alibaba.com 浙江大學 NESA Lab

Fine-Grained Fashion Similarity Prediction by Attribute-Specific Embedding Learning

Zhe Ma
NESA Lab, College of Computer Science and Technology, Zhejiang University

When Fashion MEETS IT

➢ **The rapid growth of fashion e-commerce industry.**

Computer Vision for Fashion Retrieval

➢ **Predict the similarity between two images.**

In-Domain
In-shop Clothes Retrieval

Cross-Domain
Consumer-to-Shop Clothes Retrieval

Functionality
Recommendation

Interaction
Interactive Fashion Search

Case 1: I'm not looking for the same!

The ***collar design*** of this suit is so cool. But I'd like some other types...

These search engines always return identical ones...

taobao.com

jd.com

Case 2: Which is Plagiarism?

This T-shirt has been selling like hot cakes! It's all because of its well designed *pattern*.

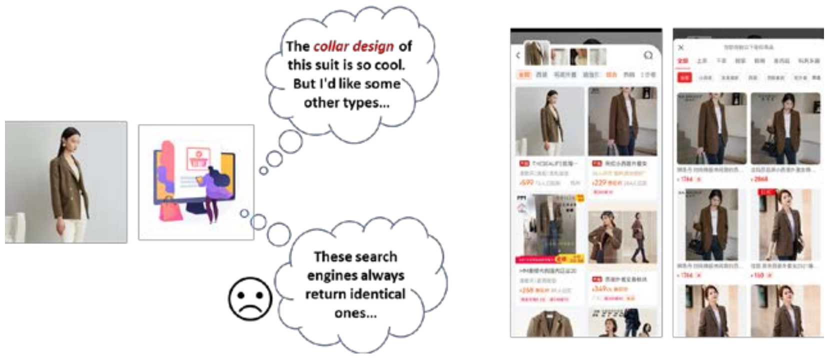
Plagiarist

Great! I *just copy a part of it* and change all the others. Then those stupid AI robots can certainly not find this!

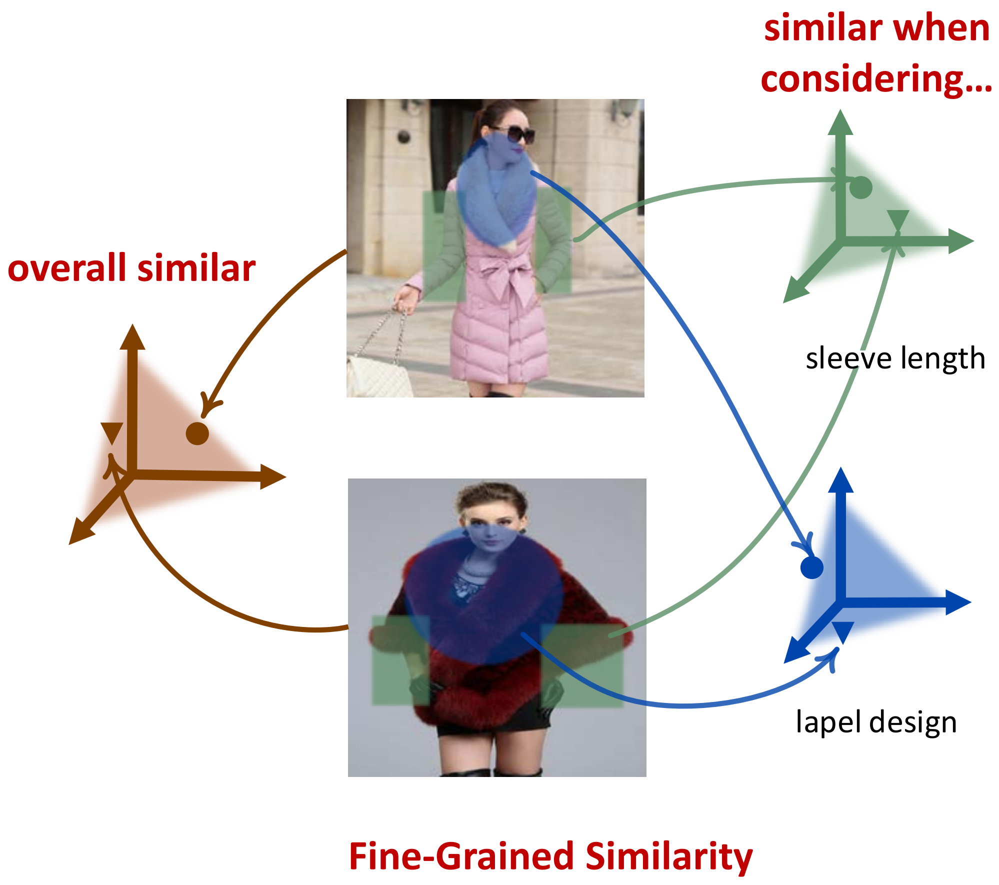
How about I change it's color. No! No! It will be easily detected!

Fine-Grained Fashion Similarity

Advanced Retrieval

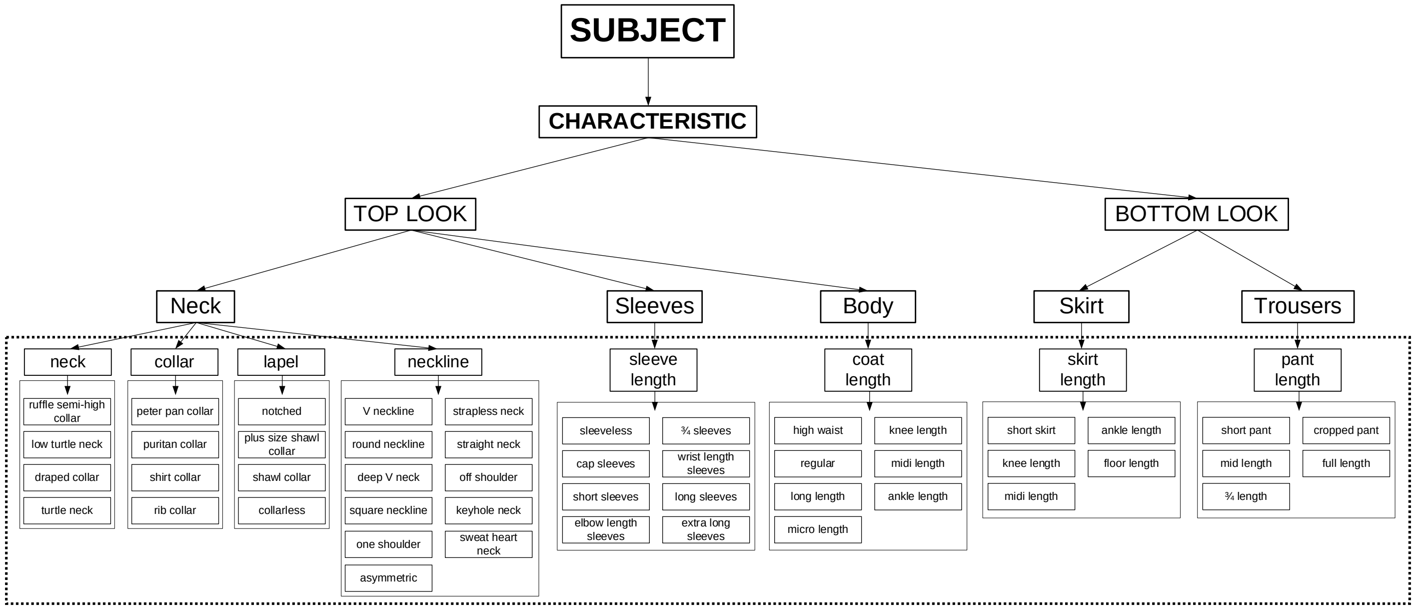


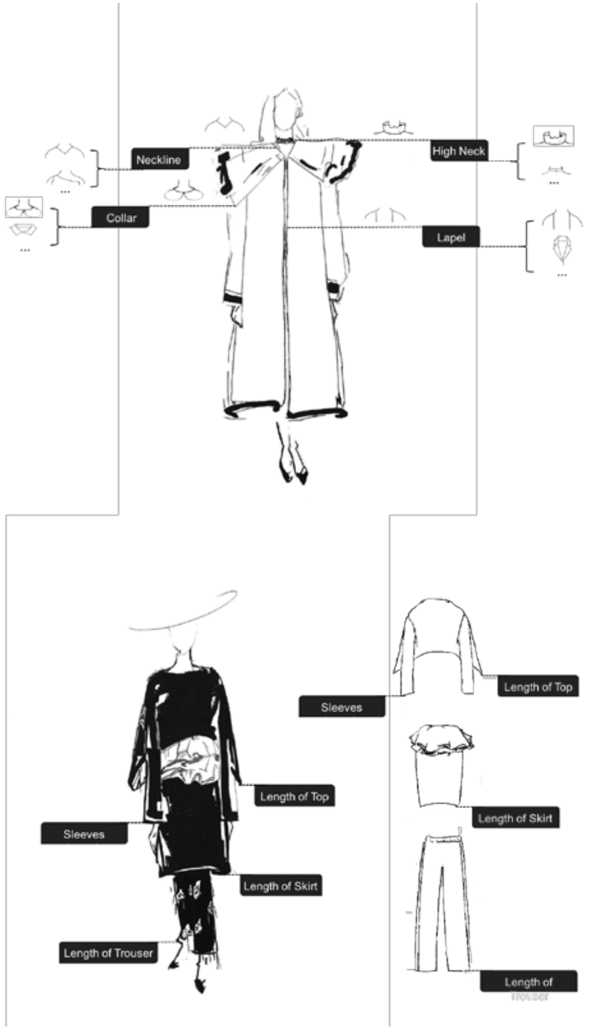
Copyright Protection



Key Problems

- ➢ **Simultaneously model multiple spaces for different fine-grained similarity notions**
- ➢ Tackle with unrestricted fashion images





- ➢ The hierarchical attribute tree can be used as well-partitioned fine-grained similarity notions of fashion items.
- ➢ Fine-grained fashion attributes are typically related to specific regions of the image and one region may corresponds to multiple attributes. Attention mechanism is a promising tool for learning different attributes.

Key Problems

- Simultaneously model multiple spaces for different fine-grained similarity notions
- **Tackle with unrestricted fashion images**

Attribute: neckline design Attribute: sleeve length Attribute: pant length

- A practical system should be capable of processing unrestricted images.
- Fashion images in practical scenario can be of high resolution(hundreds to thousands of pixels), while some attributes only correspond to minor parts of clothes.
- Typical CNN backbones take relatively low resolution(e.g., 224x224) image as input. This will lead to losing detailed information that is critical for those attributes. However, large resolution input hurts the efficiency of the model.

Our Contribution

- Conceptually, we propose to simultaneously learn multiple attribute-specific embedding spaces for fine-grained fashion similarity prediction. Each space accounts for a certain notion of similarity defined by fashion attributes.

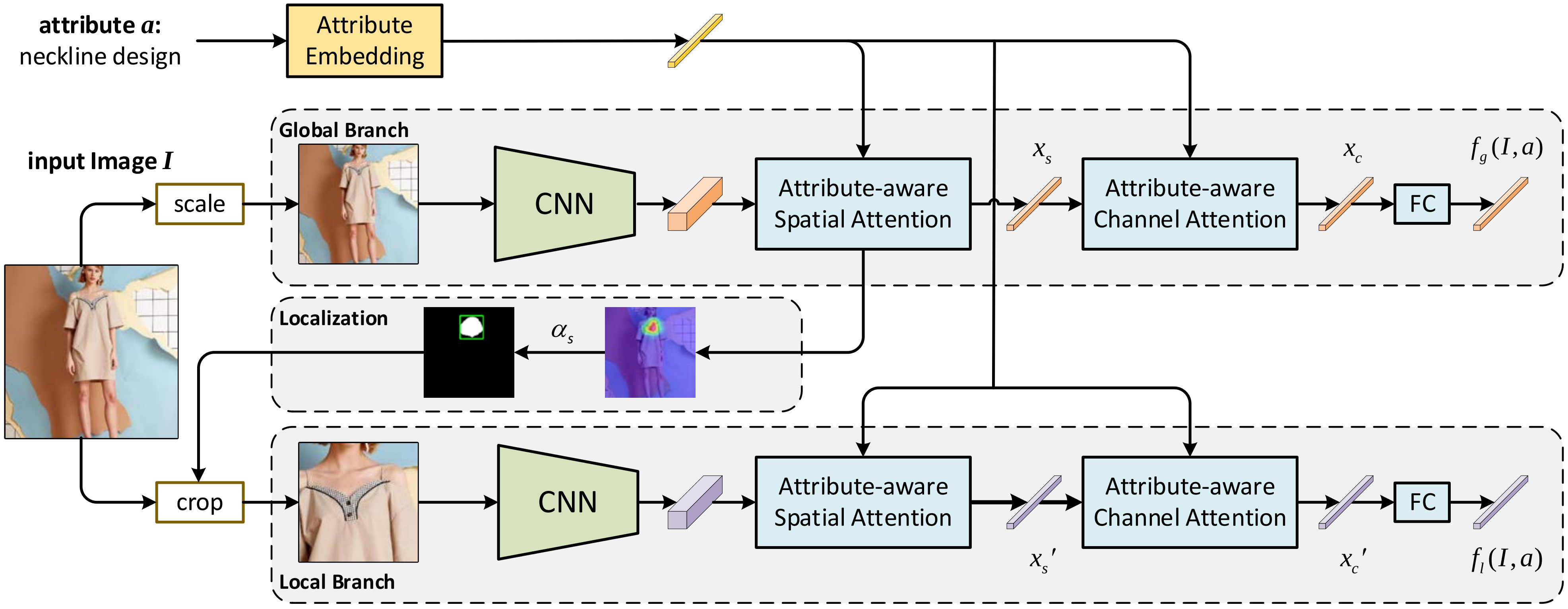
- Technically, we propose a ASEN model consisting of two branches and combined with two attention modules to learn multiple attribute-specific embedding spaces.

- Comprehensive experiments on three large-scale fashion understanding datasets demonstrate the feasibility and effectiveness of fine-grained similarity learning.

Framework

Overview

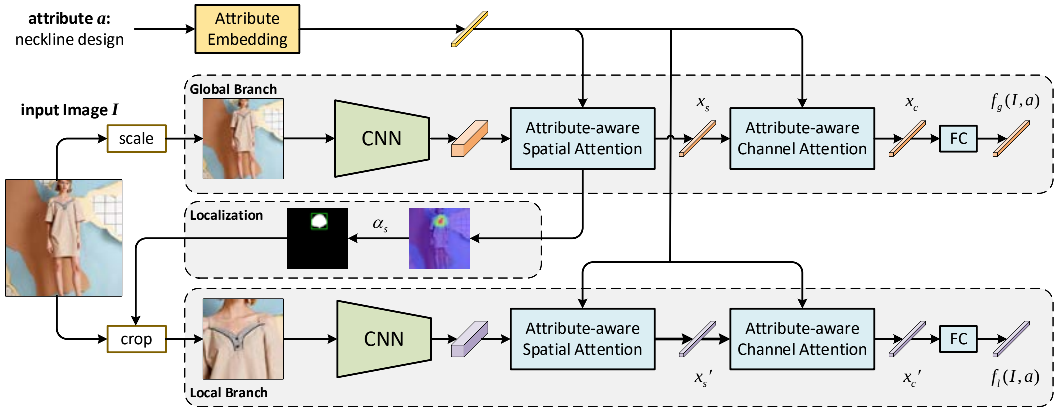
➢ Attribute-Specific Embedding Network(ASEN)

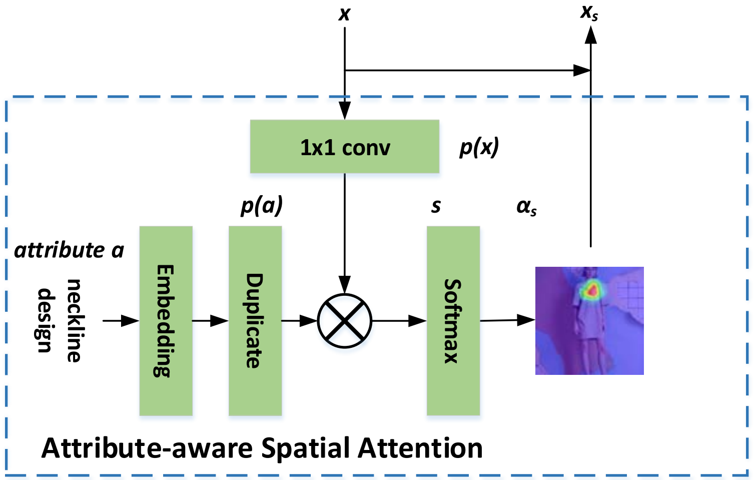


- ➢ **Attribute-aware Spatial and Attribute-aware Channel Attention**
- ➢ **Weakly-supervised Localization and Two-Branch Learning**

Attention Mechanism

➢ Attribute-aware Spatial Attention





- Fashion attributes are typically related to certain regions of clothes.
- Encode image with CNN and transform:

$$p(x) = \tanh(Conv(x))$$

- Encode attribute as embedding and duplicate along spatial dimension:

$$p(a) = \tanh(W_s a) \cdot \mathbf{1}$$

- Compare them at each location, generate spatial attention weights:

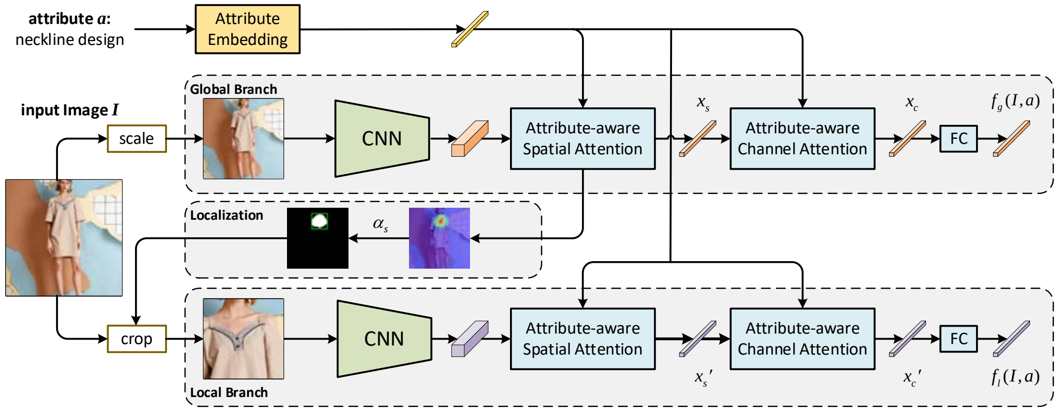
$$\alpha^s = softmax(\frac{\sum_i^c [p(a) \odot p(x)]_i}{\sqrt{(c)}})$$

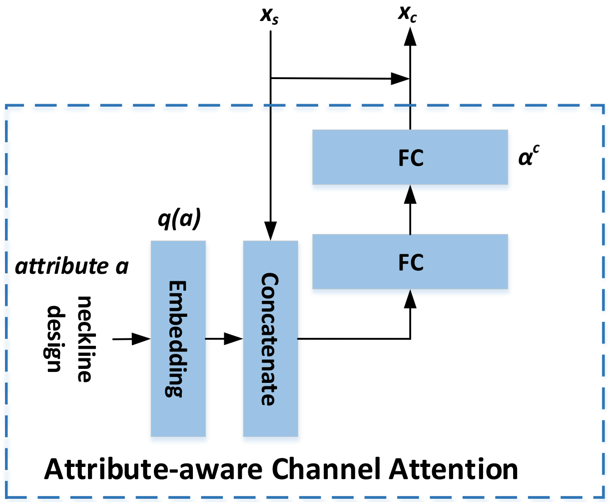
- Attend to discriminative regions:

$$x_s = \sum_j^{h \times w} \alpha_j^s x_j$$

Attention Mechanism

➢ Attribute-aware Channel Attention





- The same regions may still be related to multiple attributes, e.g., collar design and collar color.
- Encode attribute as embedding:

$$q(a) = \tanh(W_c a)$$

- Attribute-guided Squeeze-and-Excitation:

$$\alpha^c = \sigma(W_2 \delta(W_1([q(a), x_s])))$$

- Attend to discriminative channels:

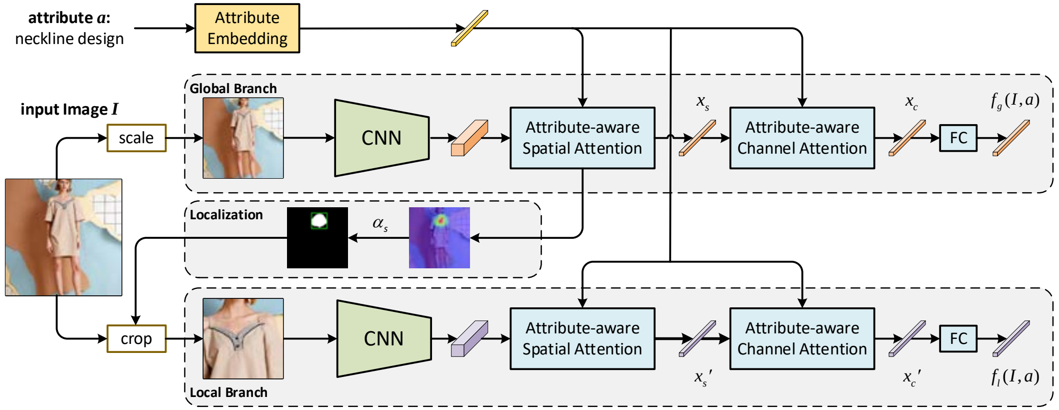
$$x_c = x_s \odot \alpha^c$$

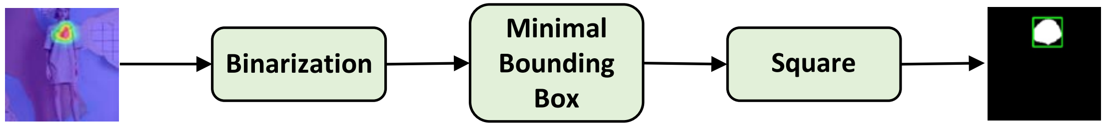
- Adjust output dimension:

$$ASEN(I, a) = W x_c + b$$

Localization and Training

Weakly-supervised Localization and Two-Branch Learning





Attention Map

Localized Patch

- Triplet ranking based metric learning:

$$\mathcal{L} = \sum \max(0, m - s(I, I^{+}|a) + s(I, I^{-}|a))$$

- Global local alignment

$$\min 1 - \frac{f_g(I,a) \cdot f_l(I,a)}{\|f_g(I,a)\|_2 \|f_l(I,a)\|_2}$$

Algorithm 1 Two-stage Training Strategy

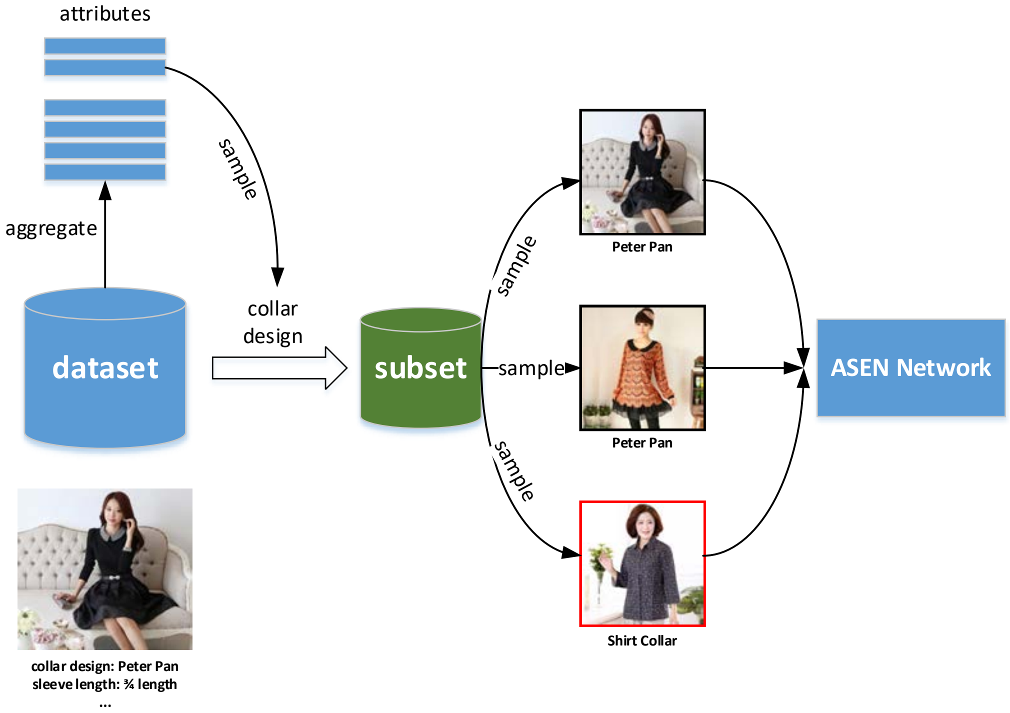
1: **input:** structure of global branch f_g and local branch f_l, triplet set $\mathcal{T}$, total training epochs E_1, E_2, batch size B, weights α, β, γ
2:
3: ▷ Stage 1: line 4-10
4: **for** $e \leftarrow 1$ **to** E_1 **do**
5:     **for** sampled minibatch $\mathcal{B} \in \mathcal{T}$ **do**
6:         calculate the global triplet ranking loss $\mathcal{L}_g$
7:         calculate gradients of the global branch $\nabla \mathcal{L}_g(\theta_g)$
8:         $\theta_g \leftarrow Adam(\nabla \mathcal{L}_g(\theta_g))$
9:     **end for**
10: **end for**
11:
12: ▷ Stage 2: line 13-25
13: **for** $e \leftarrow 1$ **to** E_2 **do**
14:     **for** sampled minibatch $\mathcal{B} \in \mathcal{T}$ **do**
15:         calculate the global triplet ranking loss $\mathcal{L}_g$
16:         obtain RoIs by the weakly-supervised localization
17:         calculate the local triplet ranking loss $\mathcal{L}_l$
18:         calculate the alignment loss $\mathcal{L}_a$
19:         $\mathcal{L} \leftarrow \alpha \mathcal{L}_g + \beta \mathcal{L}_l + \gamma \mathcal{L}_a$
20:         calculate gradients of the global branch $\nabla \mathcal{L}(\theta_g)$
21:         calculate gradients of the local branch $\nabla \mathcal{L}(\theta_l)$
22:         $\theta_g \leftarrow Adam(\nabla \mathcal{L}(\theta_g))$
23:         $\theta_l \leftarrow Adam(\nabla \mathcal{L}(\theta_l))$
24:     **end for**
25: **end for**
26:
27: **return** trained network $f_g(\cdot)$, $f_l(\cdot)$

Evaluations

Experiment Setup

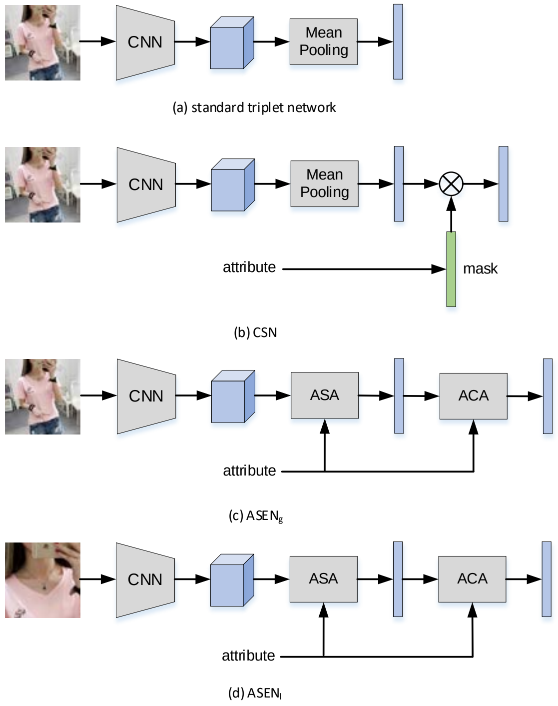
- We utilize three large-scale fashion understanding datasets, i.e., FashionAI, DARN, DeepFashion, to construct training set.
  - Aggregate appropriate fashion attributes and construct different subsets to certain attributes.
  - Random sample triplets to train our proposed ASEN.

- For evaluation, we randomly select images as queries, and other images with annotations on the same attribute as candidates.



Attribute-Specific Fashion Retrieval

➢ Comparison to baselines



FashionAI

Method	MAP for each attribute								overall MAP
	skirt length	sleeve length	coat length	pant length	collar design	lapel design	neckline design	neck design	
Random baseline	17.20	12.50	13.35	17.45	22.36	21.63	11.09	21.19	15.79
Triplet network	48.38	28.14	29.82	54.56	62.58	38.31	26.64	40.02	38.52
CSN	61.97	45.06	47.30	62.85	69.83	54.14	46.56	54.47	53.52
ASEN$_g$	64.14	54.62	51.59	65.90	71.45	66.16	60.04	60.28	60.60
ASEN$_l$	50.42	39.93	40.85	51.87	67.64	54.38	50.57	64.11	50.34
ASEN	**66.34**	**57.53**	**55.51**	**68.77**	**72.94**	**66.95**	**66.81**	**67.01**	**64.31**

DARN

Method	MAP for each attribute									overall MAP
	clothes category	clothes button	clothes color	clothes length	clothes pattern	clothes shape	collar shape	sleeve length	sleeve shape	
Random baseline	8.49	24.45	12.54	29.90	43.26	39.76	15.22	63.03	55.54	32.26
Triplet network	23.59	38.07	16.83	39.77	49.56	47.00	23.43	68.49	56.48	40.14
CSN	34.10	44.32	47.38	53.68	54.09	56.32	31.82	78.05	58.76	50.86
ASEN$_g$	38.70	48.91	52.12	58.44	54.37	58.50	36.48	82.42	59.41	54.30
ASEN$_l$	22.16	38.86	46.80	48.10	51.27	44.95	24.93	72.21	56.86	44.93
ASEN	**40.15**	**50.42**	**53.78**	**60.38**	**57.39**	**59.88**	**37.65**	**83.91**	**60.70**	**55.94**

DeepFashion

Method	MAP for each attribute					overall MAP
	texture	fabric	shape	part	style	
Random baseline	6.69	2.69	3.23	2.55	1.97	3.38
Triplet network	13.26	6.28	9.49	4.4 3	3.33	7.36
CSN	14.09	6.39	11.07	5.13	3.49	8.01
ASEN$_g$	15.01	7.32	13.32	6.27	3.85	9.14
ASEN$_l$	13.66	6.30	11.54	5.15	3.48	8.00
ASEN	**15.60**	**7.67**	**14.31**	**6.60**	**4.07**	**9.64**

Remarks

➢ ASEN outperforms all the other baseline models consistently over different datasets, different attributes.

Ablation Study

Study on attention and loss function

Method	MAP for each attribute								overall MAP
	skirt length	sleeve length	coat length	pant length	collar design	lapel design	neckline design	neck design	
ASEN_g w/o ASA	62.09	46.18	49.23	62.79	67.34	58.07	46.85	56.20	54.27
ASEN_g w/o ACA	62.84	51.46	49.07	66.08	70.36	61.47	58.14	58.02	58.53
ASEN_g	64.14	54.62	51.59	65.90	71.45	65.16	60.04	60.28	60.60
ASEN w/o $\mathcal{L}_g$	53.73	13.60	38.55	57.07	22.59	22.15	11.44	21.65	28.82
ASEN w/o $\mathcal{L}_l$	65.63	**57.78**	54.82	68.66	72.20	**67.10**	66.55	**67.56**	64.08
ASEN w/o $\mathcal{L}_a$	64.95	55.96	53.76	67.38	**74.12**	66.74	64.51	66.48	63.05
ASEN	**66.34**	57.53	**55.51**	**68.77**	72.94	66.95	**66.81**	67.01	**64.31**

Study on localization strategy

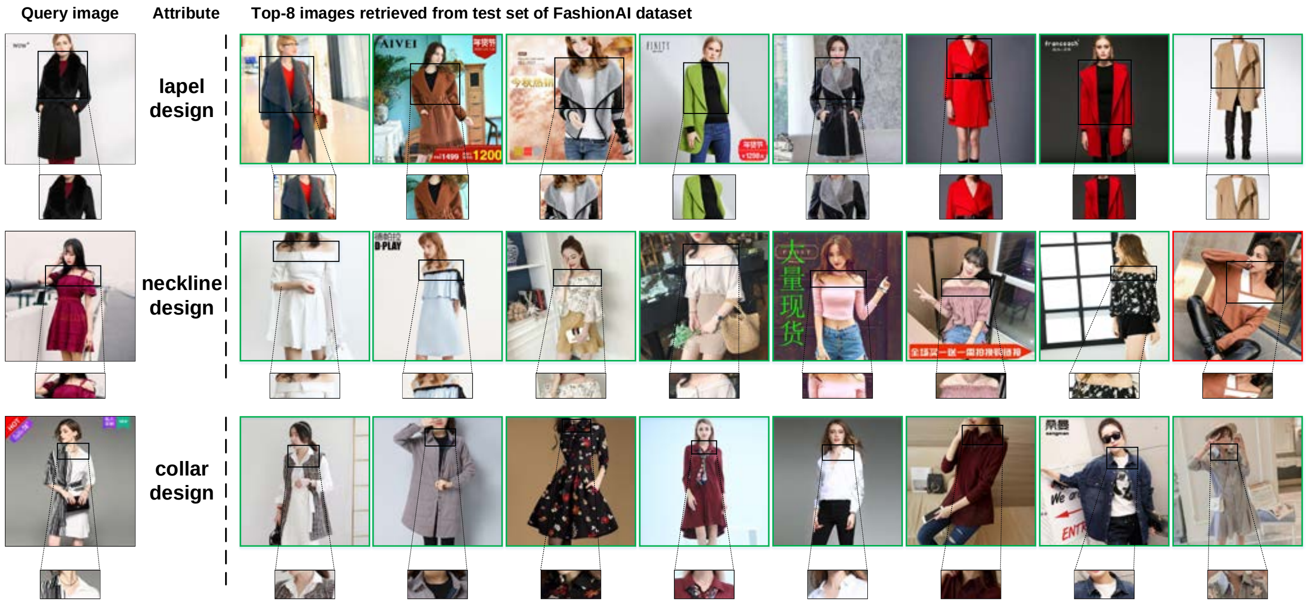
Method	MAP for each attribute								overall MAP
	skirt length	sleeve length	coat length	pant length	collar design	lapel design	neckline design	neck design	
ASEN	66.34	57.53	55.51	68.77	72.94	66.95	66.81	67.01	64.31
$\text{ASEN}_{\text{full}}$	66.51	55.43	55.37	67.61	68.58	62.30	59.09	57.45	60.81
ASEN_1	63.91	56.75	51.06	68.13	73.80	67.79	67.12	68.21	63.45
ASEN_2	65.84	58.04	54.24	68.74	72.87	66.94	66.53	67.31	64.10

Remarks

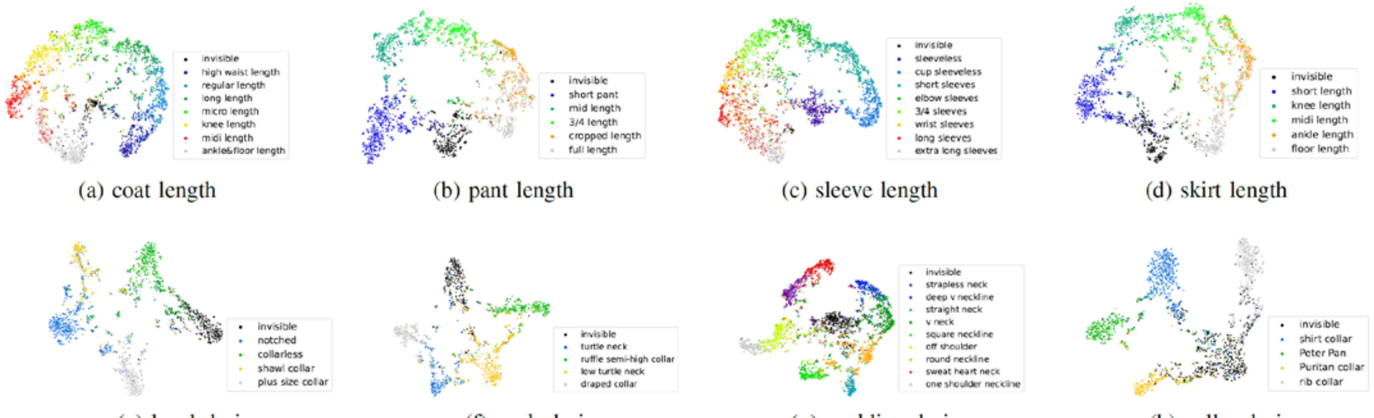
- Both attention modules contribute to learn multiple embedding spaces.
- Stable global branch is essential for two-branch training.
- ASEN is not merely ensembles. Localizing minor regions and training a local branch is much beneficial.

What has ASEN Learned?

➢ Attribute-Specific Fashion Retrieval



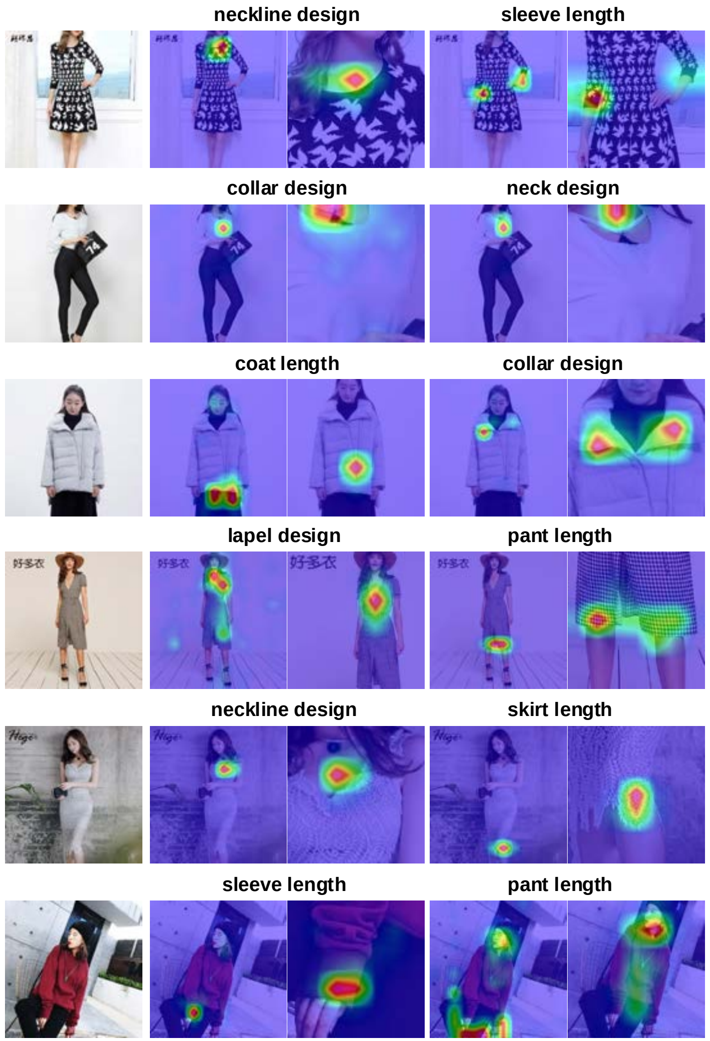
➢ t-SNE Visualization of Embedding Spaces



(a) coat length (b) pant length (c) sleeve length (d) skirt length

(e) lapel design (f) neck design (g) neckline design (h) collar design

➢ Sptial Attention Visualization



Discussion

Discussion

Conclusion

- An Attribute-Specific Embedding Network(ASEN) which learns fine-grained fashion similarity.
- Two branches extracted attribute-specific features from different perspectives.
- Two attention modules considering the locality and diversity of fashion attributes

Limitation and future work

- ASEN assumes that images come with attribute annotations.
- Automatic attribute discovery from text.

Q&A

https://github.com/maryeon/asenpp

maryeon.rs@zju.edu.cn