Self-supervised Visual Attribute Learning for Fashion Compatibility

Donghyun Kim[1], Kuniaki Saito[1], Samarth Mishra[1], Stan Sclaroff[1], Kate Saenko[1,2], Bryan A. Plummer[1]

1: Boston University, 2: MIT-IBM Watson AI Lab

2021 ICCV OCTOBER 11-17 VIRTUAL

VIPriors Workshop

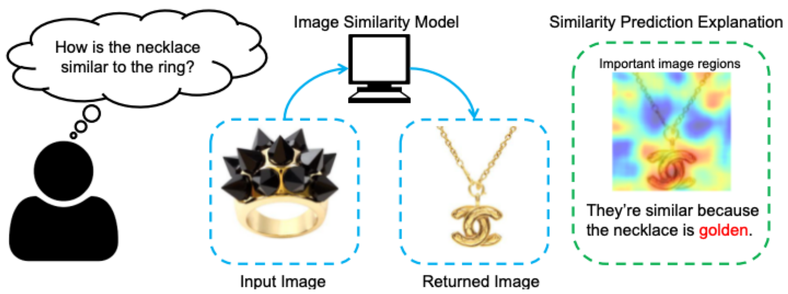
Motivation

Problem

- Self-supervised learning is very effective in visual recognition tasks, but is it still helpful for different tasks such as fashion compatibility?

(a) Object Recognition (b) Fashion Compatibility (c) Non-transferability

- Object recognition needs color invariant but shape sensitive features.
- In fashion compatibility, a system recommends fashion items compatible and complement each other when worn together in an outfit.
- Fashion compatibility needs color sensitive but shape invariant features to match different category fashion items, in which items of the same object category can be embedded far under different visual attributes.



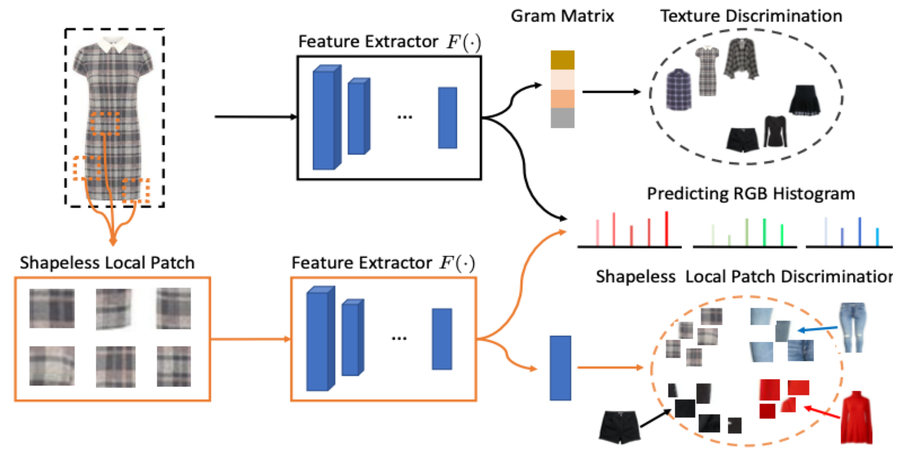
Plummer *et al.* (ECCV'20)

- Plummer *et al.* (ECCV'20) observed that similar color or texture items are likely to be compatible.
- Motivated by this, we aim to propose new self-supervised learning for fashion compatibility.

Conclusion

- While prior self-supervised learning approaches have been successful, their downstream task is mostly related to object recognition which focuses on learning object shape, which may not work different tasks.
- We explore self-supervised methods for fashion compatibility, where colors and texture are important than object shapes.

Self-supervised Tasks for Visual Attribute (S-VAL)



- We propose three self-supervised sub-tasks to learn color-sensitive but shape-invariant features:
  (1) Predicting RGB histogram
  (2) Shapeless local patch discrimination
  (3) Texture discrimination with gram matrix

Experiments

- Dataset: Polyvore Outfits, Capsule Wardrobe, Fashion-Gen
- Evaluation: Fashion compatibility and Fill in the Black (FITB) tasks
- Self-supervised Baselines: Instance Discrimination (ID), Local Aggregation (ICCV'19), Autoencoder, Colorization

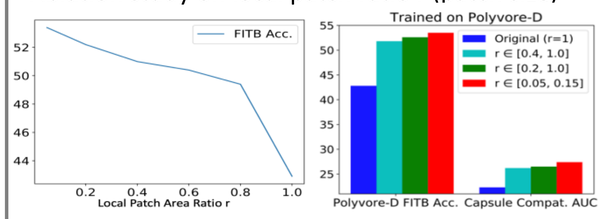
	Method	Label?	Polyvore Outfits		Capsule
			Comp. AUC	FITB acc.	Comp. AP
(a) With Label	Bi-LSTM [11]	Comp.	0.65	39.7	18.4
	SiameseNet [34]	Comp.	0.81	52.9	-
	Type-Aware Network [34]	Comp.	0.86	55.3	-
	SCE-Net [31]	Comp.	**0.91**	**61.6**	-
	Attribute Classifier	Attributes	0.73	46.3	25.0
(b) Self-sup. Baselines	ImageNet pre-trained	✗	0.66	39.1	21.1
	Capsule Network (weakly-sup.) [16]	✗	-	-	19.9
	AutoEncoder [15]	✗	0.58	34.0	19.8
	Colorization [41]	✗	0.63	34.1	18.6
	Jigsaw [26]	✗	0.52	27.9	18.6
	Rotation [9]	✗	0.53	29.4	18.5
	ID [39] w/ color distortion	✗	0.57	30.8	18.9
	ID [39] w/o color distortion	✗	0.74	45.9	23.3
	LA [42] w/ color distortion	✗	0.56	30.4	19.1
	LA [42] w/o color distortion	✗	0.74	46.3	24.0
(c) S-VAL (Ours)	Predicting RGB histogram (RGB)	✗	0.77	47.2	23.3
	Shapeless Local Patch Disc. (SLPD)	✗	0.83	54.6	27.7
	Texture Disc (TD)	✗	0.77	50.3	25.2
	RGB + SLPD	✗	0.83	55.4	27.7
	RGB + SLPD + TD	✗	**0.84**	**55.8**	**27.9**

Acknowledgement

This work was supported by NSF and the DARPA LwLL program.

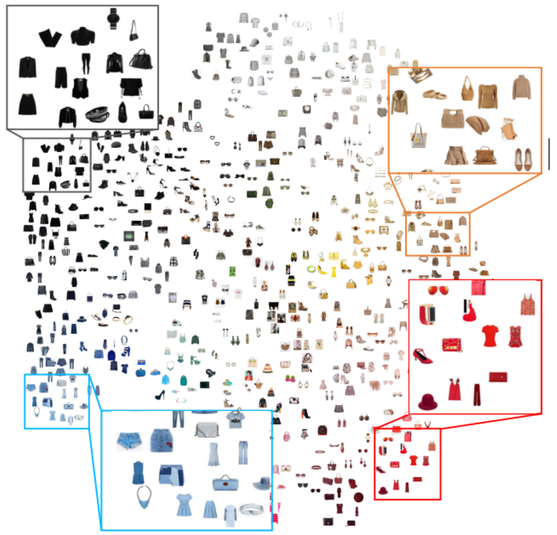
Experiments

Ablation study on local patch ratio r (patch size)

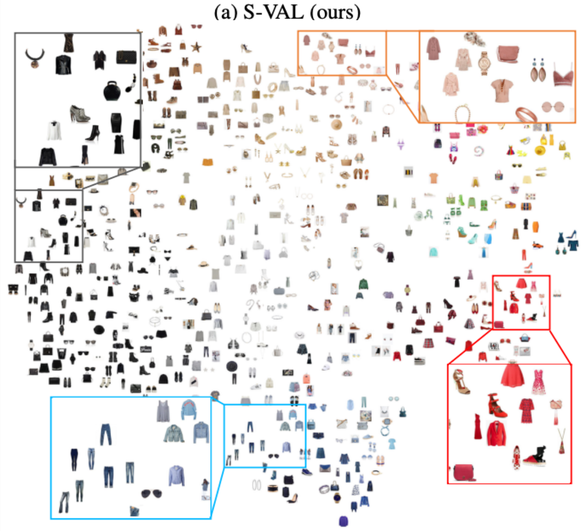


- Using Low area ratio r improves performances.

Feature visualizations



(a) S-VAL (ours)



(b) Siamese Network (supervised)

Similar to (b) the supervised model, (a) our unsupervised model, learns a similar embedding which embeds items with similar visual attributes (e.g., colors and texture)