哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

FashionSAP: Symbols and Attributes Prompt for Fine-grained Fashion Vision-Language Pre-training

HanYunpeng, ZhangLisai, ChenQingcai, ChenZhijian, Li Zhonghua, YangJianxin, CaoZhao

Poster Session WED-PM-255

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Introduction

- **Background**

Vision-Language Pre-training(VLP)

(a) General item

Caption: *a young man in a suit securing his tie.*

(b) Fashion item

Caption: *long sleeve shirt in red, white, and black plaid, single-button barrel cuffs, ...*

Attribute(b): ***Season****: spring-summer;* ***Gender****: men, ...*

- **Fashion concept with knowledge in fine-grained**
- **Fine-grained fashion features in attribute-level**

Methodology

- **Preliminary**

Fashion Symbols:

Fashion Symbols	Categories	Definition Rules
TOPS	tops, shirt, polo, sweater, ...	upper body
DRESSES	dress, suit, shift, ...	up-to-lower body
SKIRTS	skirt, sarong, slit, kilt, ...	lower body
COATS	jacket, parka, blazer, duffle, ...	associated with others
PANTS	jeans, shorts, breeches, ...	lower body
SHOES	boots, sneakers, pump, loafers, ...	feet
BAGS	clutches, pouches, wristlet, ...	decorative
ACCESSORIES	ring, sunglasses, accessories, …	decorative
OTHERS	swim-wear, lingerie, lounge-wear, ...	-

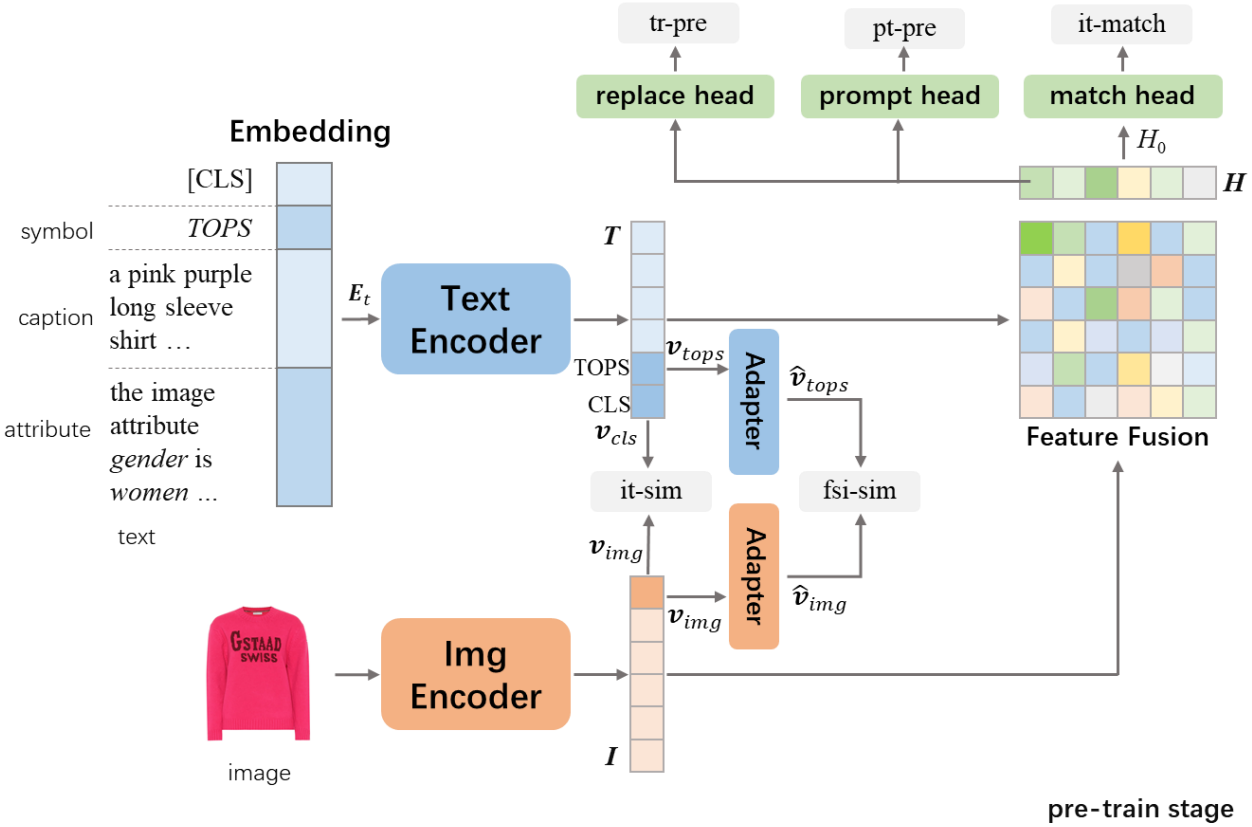
Attributes prompt templates:

- **Enumerable Attribute** : {gender: men, season: spring-summer}
  Template: `the image attribute [Attr_name] is [Attr_value]`

- **Binary Attribute** : {red, spring-summer, men}
  Template: `is image attribute [Attr_value] ? [Yes_or_No]`

Methodology

- **Architecture**



Introduction

- **Background**

(a) General item

Caption: *a young man in a suit securing his tie.*

(b) Fashion item

Caption: *long sleeve shirt in red, white, and black plaid, single-button barrel cuffs, ...*

Attribute(b): ***Season****: spring-summer;* ***Gender****: men, ...*

- **Vision-Language Pre-training(VLP)**

  VLP performances significantly in many tasks by aiming at learning multimodal knowledge from aligning the object from both vision and language.

- **Fine-grained fashion concepts -> Fashion Symbols**

  Fashion concepts can be summarized to specific representations with human knowledge.

- **Fine-grained fashion features -> Attributes Prompt**

  Fine-grained features can be represented in attribute-level explicitly.

Methodology

- **Preliminary**

Fashion Symbols	Categories	Definition Rules
TOPS	tops, shirt, polo, sweater, ...	upper body
DRESSES	dress, suit, shift, ...	up-to-lower body
SKIRTS	skirt, sarong, slit, kilt, ...	lower body
COATS	jacket, parka, blazer, duffle, ...	associated with others
PANTS	jeans, shorts, breeches, ...	lower body
SHOES	boots, sneakers, pump, loafers, ...	feet
BAGS	clutches, pouches, wristlet, ...	decorative
ACCESSORIES	ring, sunglasses, accessories, …	decorative
OTHERS	swim-wear, lingerie, lounge-wear, ...	-

All fashion categories are mapped to 9 predefined symbols by experts following rules:

- **Body Part**: fashion items that are associated with a specific part of the human body.
- **Function**: fashion items that are optionally used for decoration and can be dressed on multiple body parts.

Methodology

Preliminary

Two templates are designed to cover two presentation formats of attributes:

- **Enumerable Attribute** :

  Sample: {gender : men, season : spring-summer}

  Template: *the image attribute [Attr_name] is [Attr_value]*

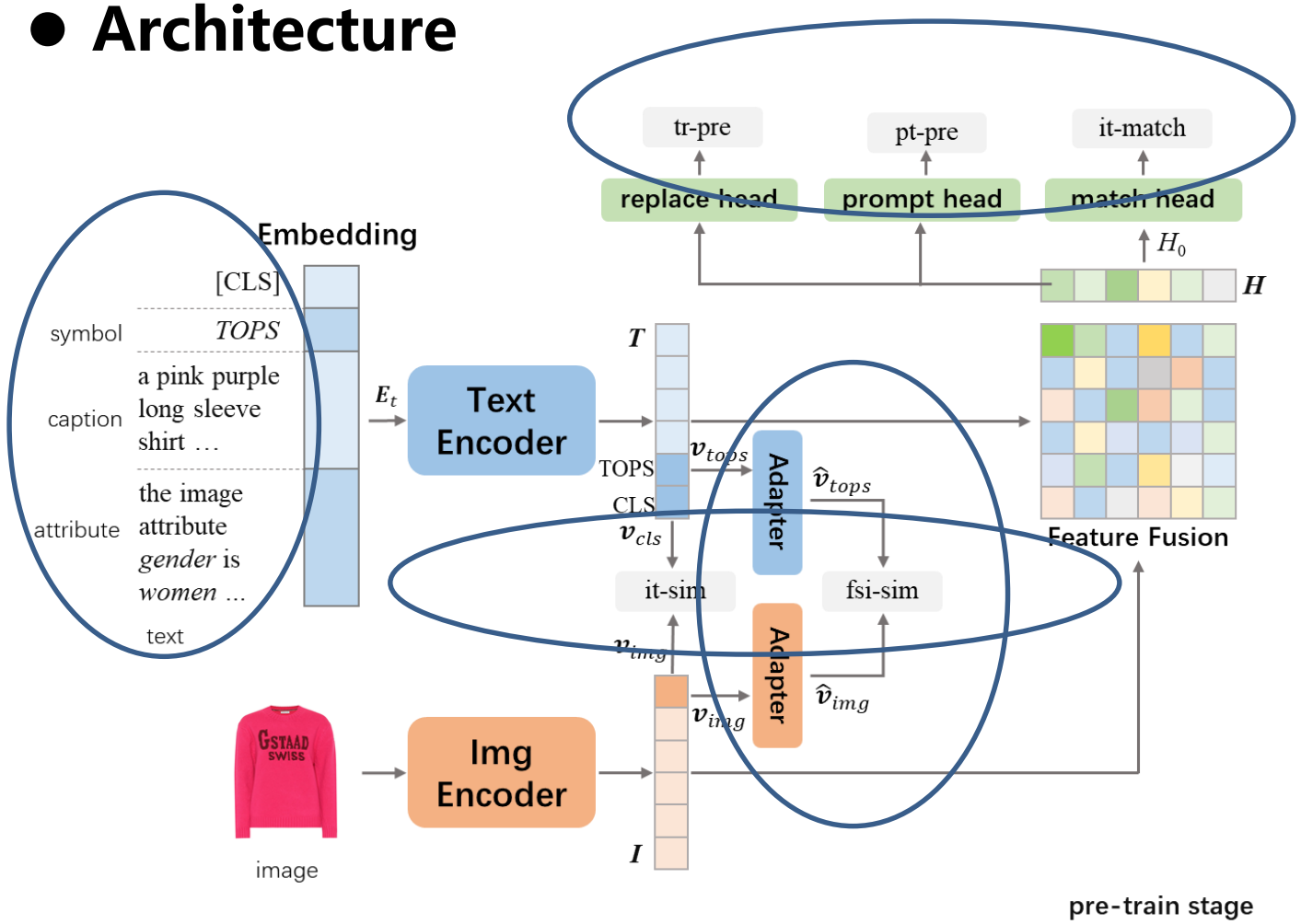
- **Binary Attribute** :

  Sample: {red, spring-summer, men}

  Template: *is image attribute [Attr_value] ? [Yes_or_No]*

Methodology

- **Architecture**



•*Pretrain-finetune* paradigm.

•Fashion symbols, caption and prompt attributes are concatenated into integration text and are fed to Embedding layer.

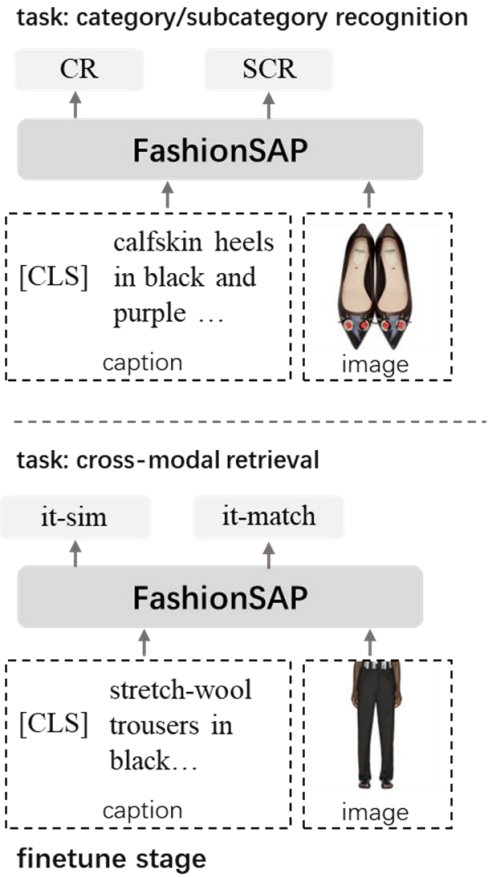
•Fashion symbols features are adapted to close to the adapted image feature.

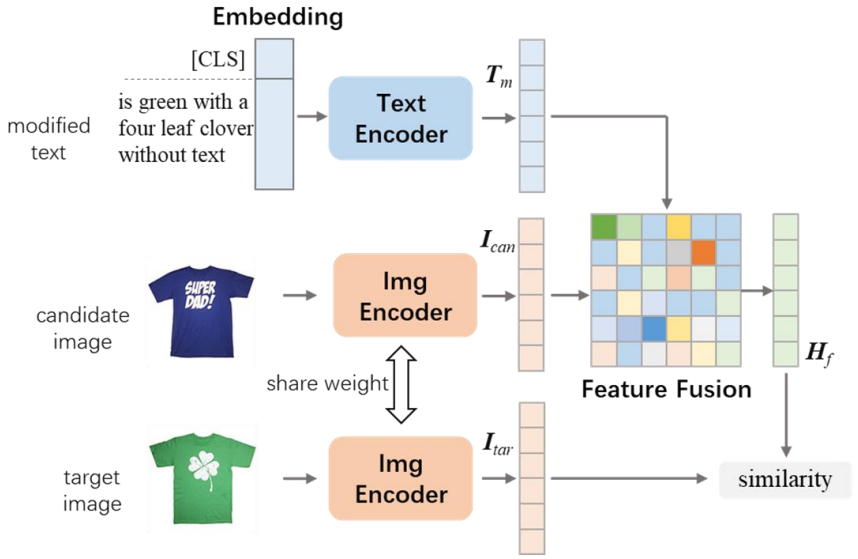
•Multi-task are performed in pre-training stage and the downstream task are optimized only by target task.

$$\mathcal{L}_{fsis} = \frac{1}{B}[1 - \sum_{b=1}^{B} \frac{1}{2}[\hat{v}^b_{img}(\hat{v}^b_{symbol})^\top + 1]]$$

Methodology

- **Downstream Tasks**





- **Category/Subcategory recognition**
- **Cross-modal retrieval (I2T,T2I)**
- **Text modified image retrieval**

Experiments

- **Comparison**

Methods	I2T			T2I			Mean
	R@1	R@5	R@10	R@1	R@5	R@10	R@1
VL-BERT [34]	19.26	39.90	46.05	22.63	36.48	48.52	20.95
ViLBERT [25]	20.97	40.49	48.21	21.12	37.23	50.11	21.05
Image-BERT [29]	22.76	41.89	50.77	24.78	45.20	55.90	23.77
OSCAR [18]	23.39	44.67	52.55	25.10	49.14	56.68	24.25
FashionBERT [4]	23.96	46.31	52.12	26.75	46.48	55.74	25.36
KaleidoBERT [49]	27.99	60.09	68.37	33.88	60.60	68.59	30.94
EI-CLIP [26]	38.70	72.20	84.25	40.06	71.99	82.90	39.38
CommerceMM [45]	41.60	64.00	72.80	39.60	61.50	72.70	62.75
ALBEF [17]	63.97	88.92	94.41	60.52	84.99	91.45	62.20
FashionViL [7]	65.54	91.34	96.30	61.88	87.32	93.22	63.71
FashionSAP(Resnet50)	67.23	91.30	96.41	64.11	88.24	94.31	65.67
FashionSAP(ViT-B16)	71.14	92.21	96.52	69.07	89.81	94.75	70.11
FashionSAP	**73.14**	**92.80**	**96.87**	**70.12**	**91.76**	**96.38**	**71.63**

Table 2. Cross-modal retrieval result on FashionGen [31] in the sub set of evaluation following previous work.

Methods	CR		SCR	
	Acc	Macro-F	Acc	Macro-F
F-BERT [4]	91.25	70.50	85.27	62.00
K-BERT [49]	95.07	71.40	88.07	63.60
F-ViL [7]	97.48	88.60	92.23	83.02
FashionSAP	**98.34**	**89.84**	**94.33**	**87.67**

Table 5. CR and SCR results on FashionGen [31].

Experiments

- **Comparison**

Methods	Dress		Toptee		Shirt		Mean	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
CIRR [22]	17.45	40.41	21.64	45.38	17.53	38.81	18.87	41.53
VAL [1]	22.53	44.00	27.53	51.68	22.38	44.15	24.15	46.61
CosMo [13]	25.64	50.30	29.21	57.46	24.90	49.18	26.58	52.31
DCNet [12]	28.95	56.7	30.44	58.29	23.95	47.3	27.78	54.10
FashionVLP [5]	32.42	60.29	38.51	68.79	31.89	58.44	34.27	62.51
FashionViL [7]	33.47	59.94	34.98	60.79	25.17	50.39	31.21	57.04
FashionSAP	**33.71**	**60.43**	**41.91**	**70.93**	**33.17**	**61.33**	**36.26**	**64.23**

Table 4. Text modified image retrieval performance in FashionIQ [40]

Experiments

- **Ablation**

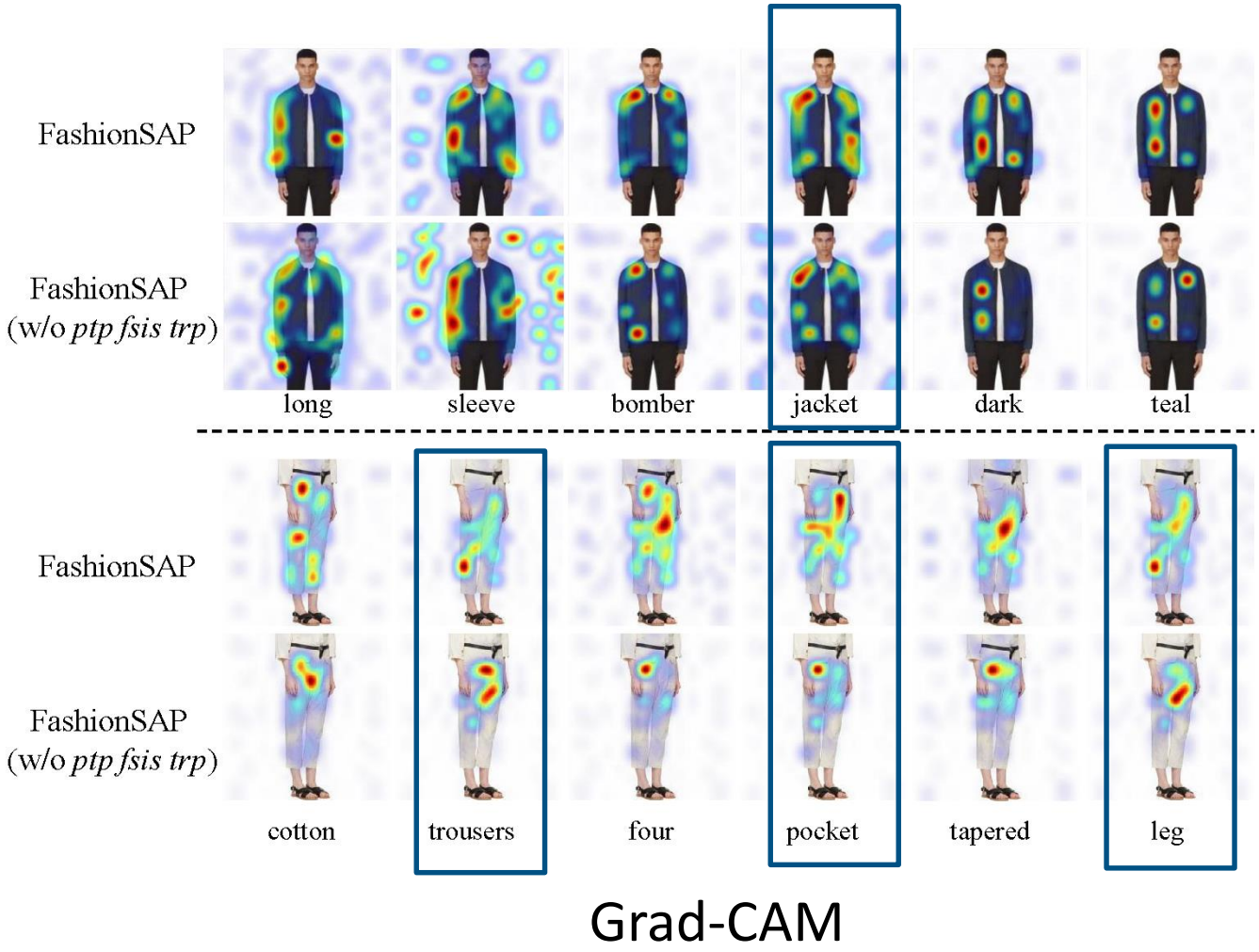
ptp	*trp*	*fsis*	I2T R@1	T2I R@1	CR Macro-F	SCR Macro-F	TMIR R@10
			43.84	53.24	84.50	84.42	30.02
✓			51.99	53.78	86.32	86.03	34.40
✓	✓		52.09	55.54	86.51	86.65	35.01
✓	✓	✓	**54.43**	**62.82**	**89.84**	**87.67**	**36.26**

Table 6. Ablation study results for proposed tasks(*ptp*, *fsis*, *trp*) on five downstream tasks.

- *ptp* task brings an improvement for I2T. As more fine-grained attributes are encoded into text side with prompt.
- *fsis* task brings an improvement for T2I. As the fashion symbol can capture the information form the text to the image.

Experiments

- **Visualization**



Grad-CAM

- FashionSAP can pay proper attention to the whole region of the object (`trousers`, `leg`, `jacket`) rather than the sub-region.
- FashionSAP can also find all positions of `pockets` in the attention maps rather than only one.

Conclusion

- Fine-grained fashion concepts and attribute-level features benefits to VLP training.
- Diversified prompt will be helpful in fashion task as what the Large Language Model shows.
- More fine-grained symbols can be used to include more fashion item.

Thank you!

Code

Paper