FashionViL: Fashion-Focused Vision-and-Language Representation Learning

Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, Tao Xiang

xiao.han@surrey.ac.uk https://github.com/BrandonHanx/mmf

Introduction

Background: A powerful fashion product search system can improve product discoverability, accessibility, buyer and seller engagement, and conversion rates in e-commerce.
Motivation: Existing V+L methods are inadequate for fashion domain as they overlook the unique characteristics of both the fashion V+L data (fine-grained + multiple images) and downstream tasks (more flexible and diverse).
Contributions: (1) A novel V+L pre-training framework with two fashion-tailored pretext tasks (Multi-View Contrastive Learning & Pseudo-Attributes Classification); (2) A flexible architecture design with a shared text encoder and fusion encoder, which can be easily adapted to diverse fashion tasks; (3) SOTA performance on 5 fashion tasks.

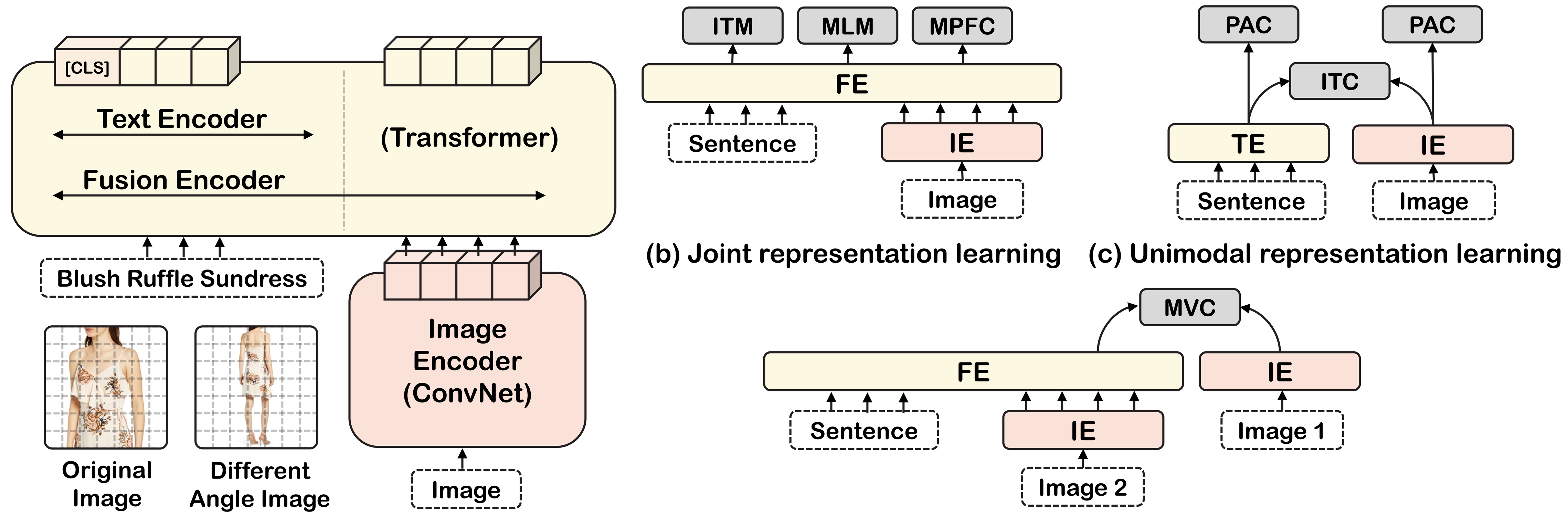
Title: Strappy floral tiered maxi dress

Style: Ivory sunrise

Description: Sun baked flower fall around the tiered skirt of a romantic maxi dress fashioned with ruffled trim at the neckline and an adjustable tie belt at the waist.

Caption: A man is standing in front of a brick storefront wearing a black jacket.

Architectures



(a) Architecture overview (b) Joint representation learning (c) Unimodal representation learning (d) Multi-view multimodal representation learning

Learning Objectives

Multi-view contrastive learning (MVC)
pulling closer the visual representation of one image to the fused multimodal representation of another image+text

$$\mathcal{L}_{\text{MVC}} = \frac{1}{2}\left[\mathcal{L}_{\text{InfoNCE}}([\mathbf{w};\mathbf{d}],\mathbf{v}) + \mathcal{L}_{\text{InfoNCE}}(\mathbf{v},[\mathbf{w};\mathbf{d}])\right]$$

Pseudo-attribute classification (PAC)
predicting pseudo-attributes extracted from fashion corpus

$$\mathcal{L}_{\text{PAC}} = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D}\mathbb{E}_{a\sim A}\left[a \log P_\theta(a|\mathbf{w}) + a \log P_\theta(a|\mathbf{v})\right]$$

Masked patch feature classification (MPFC)
predicting patch labels generated by pre-trained VQVAE

$$\mathcal{L}_{\text{MPFC}} = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D} \log P_\theta\left(\mathbf{v}_{\mathbf{m}}^{\mathbf{t}}|\mathbf{v}_{\backslash\mathbf{m}},\mathbf{w}\right)$$

Image-text contrastive learning (ITC)
pulling closer the visual representation and textual representation in a CLIP-like manner

$$\mathcal{L}_{\text{ITC}} = \frac{1}{2}\left[\mathcal{L}_{\text{InfoNCE}}(\mathbf{w},\mathbf{v}) + \mathcal{L}_{\text{InfoNCE}}(\mathbf{v},\mathbf{w})\right]$$

Masked language modeling (MLM)
predicting masked words in a BERT-like manner

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D} \log P_\theta\left(\mathbf{w}_{\mathbf{m}}|\mathbf{w}_{\backslash\mathbf{m}},\mathbf{v}\right)$$

Image-text matching (ITM)
verifying input pair in an ALBEF-like manner

$$\mathcal{L}_{\text{ITM}} = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim H} \log P_\theta\left(z|\mathbf{w},\mathbf{v}\right)$$

Main Results

Cross-modal retrieval on FashionGen

Methods		VSE++	ViLBERT	VLBERT	Image-BERT	Fashion-BERT	OSCAR	Kaleido-BERT	Ours -e2e -pt	Ours -pt	Ours
ITR	R@1	4.59	20.97	19.26	22.76	23.96	23.39	27.99	21.13	58.84	**65.54**
	R@5	14.99	40.49	39.90	41.89	46.31	44.67	60.09	46.82	89.46	**91.34**
	R@10	24.10	48.21	46.05	50.77	52.12	52.55	68.37	58.71	95.84	**96.30**
TIR	R@1	4.60	21.12	22.63	24.78	26.75	25.10	33.88	25.83	57.16	**61.88**
	R@5	16.89	37.23	36.48	45.20	46.48	49.14	60.60	51.54	84.34	**87.32**
	R@10	28.99	50.11	48.52	55.90	55.74	56.68	68.59	63.53	91.90	**93.22**
Mean		15.69	36.36	35.47	40.22	41.89	41.92	53.25	44.59	79.59	**82.60**

Text-guided image retrieval on FashionIQ

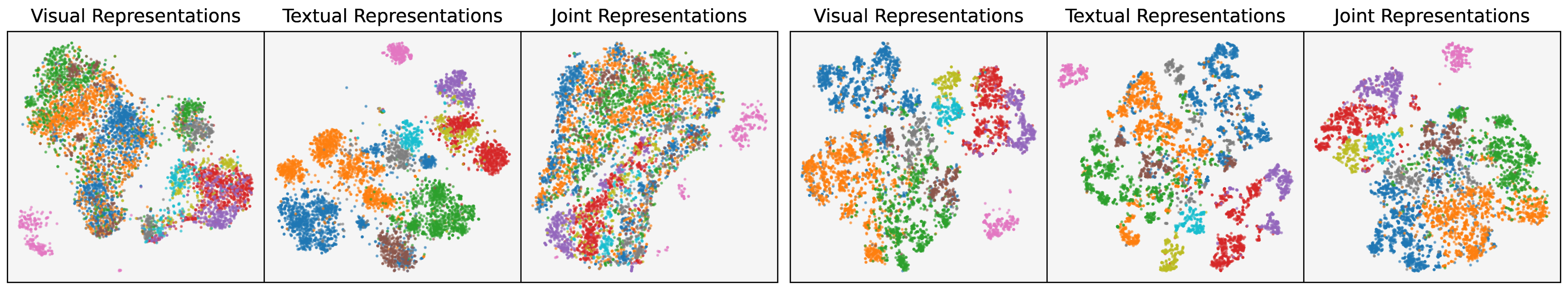
Image Encoder		Fixed ResNet 152				ResNet 50					
Fusion Module / Text Encoder		CIRR-*pt*	CIRR	Ours-*pt*	Ours	TIRG GRU	VAL GRU	CoSMo GRU	TIRG BERT	Ours-*pt*	Ours
Dress	R@10	14.38	17.45	20.97	**22.66**	23.65	26.28	24.49	27.17	28.46	**33.47**
	R@50	34.66	40.41	42.64	**46.60**	49.93	50.25	51.01	53.25	54.24	**59.94**
Shirt	R@10	13.64	17.53	17.62	**18.74**	21.98	21.69	18.99	22.28	22.33	**25.17**
	R@50	33.56	38.31	41.32	**41.56**	46.61	45.53	43.57	45.58	46.07	**50.39**
Toptee	R@10	16.44	21.64	21.67	**25.29**	27.84	27.43	25.19	27.84	29.02	**34.98**
	R@50	38.34	45.38	46.46	**50.28**	55.07	56.25	54.00	57.11	57.93	**60.79**
Mean		25.17	30.20	31.78	**34.19**	37.51	37.91	36.21	38.87	39.67	**44.12**

(Sub)category recognition on FashionGen

Methods		Fashion-BERT	OSCAR	Kaleido-BERT	Ours -pt	Ours
CR	Acc	91.25	91.79	95.07	97.07	**97.48**
	Macro$\mathcal{F}$	70.50	72.70	71.40	84.72	**88.60**
SCR	Acc	85.27	84.23	88.07	91.45	**92.23**
	Macro$\mathcal{F}$	62.00	59.10	63.60	78.13	**83.02**
Mean		77.76	76.96	79.54	87.84	**90.33**

Outfit complementary item retrieval on Polyvore

Methods		CSA-Net	ADDE-O	CSA-Net *reproduced*	Ours -*pt*	Ours
OCIR	R@10	5.93	6.18	2.69	4.38	**5.83**
	R@30	12.31	13.79	6.29	10.54	**12.61**
	R@50	17.85	18.60	9.14	14.77	**17.49**
Mean		12.03	12.86	6.04	9.90	**11.98**



(a) Without multimodal pre-training (b) With all pre-training tasks

Ablation Study

Pre-training Tasks	ITR	TIR	TGIR	SCR	OCIR	Meta-sum
None	62.50	68.09	39.67	84.79	9.90	265.04
MVC (use augmented image only)	62.85	68.58	40.50	84.86	9.53	266.32
MPFC	62.10	68.12	40.22	86.39	10.05	266.88
MLM (mask attribute words only)	62.32	67.93	40.46	85.83	10.38	266.92
MLM	62.15	67.43	40.29	86.72	10.38	266.97
PAC	63.15	69.30	40.68	86.36	9.58	269.07
MVC	63.30	68.32	40.94	85.99	10.83	269.38
ITC	64.63	70.61	43.13	86.25	10.69	275.31
ITC + MLM + MPFC	64.28	70.02	43.31	87.21	11.12	275.94
ITC + MLM + MPFC + ITM	64.37	70.44	43.56	87.17	11.08	276.62
ITC + MLM + MPFC + ITM + MVC	64.88	70.34	43.94	87.12	11.56	277.84
ITC + MLM + MPFC + ITM + MVC + PAC	65.00	70.63	44.12	87.63	11.98	279.36
same as (11) but w/o sharing TE and FE	64.16	69.15	42.87	86.22	11.31	273.71