FAME-ViL: Multi-Tasking Vision-Language Model for Heterogeneous Fashion Tasks

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

xiao.han@surrey.ac.uk https://github.com/BrandonHanx/FAME-ViL

Highlight

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

Background: Various real-world multi-modal Vision-and-Language (V+L) tasks, including recognition, retrieval, and image captioning, exist in the fashion domain, with applications in e-commerce to enhance product discoverability, engagement, and customer conversion rates, but they are *heterogeneous* in terms of *input/output formats* and *dataset sizes*.

Motivation: Existing methods tackle these fashion tasks independently, leading to *parameter redundancy* and a *lack of inter-task relatedness*.

Contributions: (1) A task-versatile pipeline with *two novel adapters*, adapting a pre-trained CLIP model for all tasks; (2) An efficient and effective *multi-task training strategy* supporting heterogeneous task modes in one unified model; (3) *SOTA performance* on 5 fashion tasks with *61.5% parameter saving*.

Cross-Modal Retrieval (XMR)

Text Query: **Long sleeve relaxed-fit silk blazer in light peach. Shawl collar. Single-button closure and patch pockets at front. Breast pocket. Slits at sleeve cuffs. Vented at back.**

Text-Guided Image Retrieval (TGIR)

Reference Image — Modifying Text: **is a black and white dress, is strapless**

Sub-Category Recognition (SCR)

Slouchy lamb nubuck patrol hat in black. Wrinkling and light distressing throughout. Fully lined.

Predicted Class: `[FLAT CAPS]`

Fashion Image Captioning (FIC)

Generated Caption: **Grey & brown camo print tank top. Relaxed-fit tank top in tones of grey, brown, and black. Signature snake graphic print throughout. Ribbed crewneck collar. Tonal stitching.**

Learning Objectives

$$\mathcal{L}_{\text{XMR}} = \frac{1}{2}\left[\mathcal{L}_{\text{InfoNCE}}(\mathbf{T},\mathbf{I}) + \mathcal{L}_{\text{InfoNCE}}(\mathbf{I},\mathbf{T})\right]$$

$$\mathcal{L}^{\text{D}}_{\text{XMR}} = \frac{1}{2B}\sum_{b}^{B}\left(\text{KL}\left(\mathbf{s}_{b,\cdot} \parallel \tilde{\mathbf{s}}_{b,\cdot}\right) + \text{KL}\left(\mathbf{s}_{\cdot,b} \parallel \tilde{\mathbf{s}}_{\cdot,b}\right)\right)$$

$$\mathcal{L}_{\text{SCR}} = -\mathbb{E}_{(I,T)\sim D}\log P\left(f_\theta^{[f]}(I,T)\right), \mathcal{L}^{\text{D}}_{\text{SCR}} = \text{KL}\left(f_\theta^{[f]}(I,T) \parallel g_{\text{scr}}(I,T)\right)$$

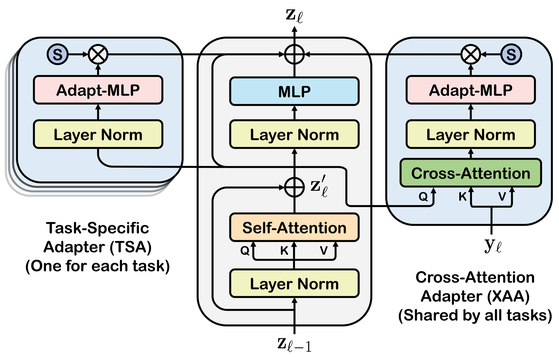
$$\mathcal{L}_{\text{TGIR}} = \mathcal{L}_{\text{InfoNCE}}\left((\mathbf{I}^{\text{r}},\mathbf{T}),\mathbf{I}^{\text{t}}\right), \mathcal{L}^{\text{D}}_{\text{TGIR}} = \frac{1}{B}\sum_{b}^{B}\text{KL}\left(\mathbf{s}_{(b,b),\cdot} \parallel \tilde{\mathbf{s}}_{(b,b),\cdot}\right)$$

$$\mathcal{L}_{\text{FIC}} = -\mathbb{E}_{(I,T)\sim D}\sum_{a=1}^{A}\log P\left(T_a \middle| f_\theta^{[g]}(I;T_{<a})\right)$$

$$\mathcal{L}^{\text{D}}_{\text{FIC}} = \sum_{a=1}^{A}\text{KL}\left(f_\theta^{[g]}(I;T_{<a})_a \parallel g_{\text{fic}}(I;T_{<a})_a\right)$$

$$\mathcal{L} = \mathcal{L}_{[\text{task}]} + \mathcal{L}^{\text{D}}_{[\text{task}]}, \quad [\text{task}] \in \{\text{XMR},\text{TGIR},\text{SCR},\text{FIC}\}$$

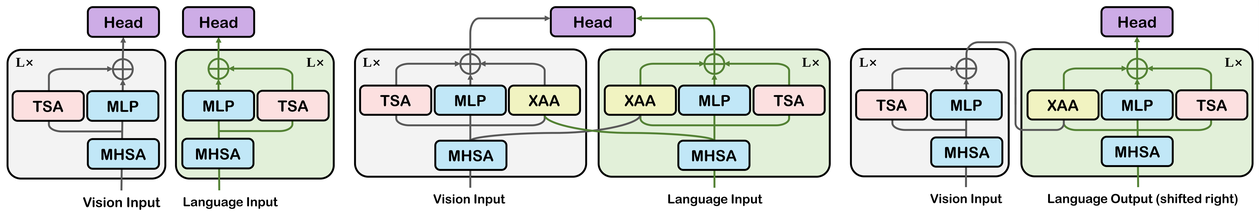
Task-Versatile Transformer Layer



Task-Specific Adapter (TSA) (One for each task)

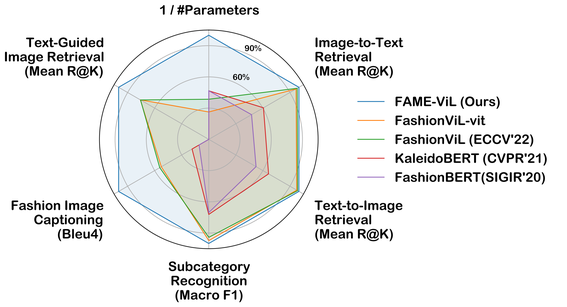
Cross-Attention Adapter (XAA) (Shared by all tasks)

Overview of Three Operational Modes



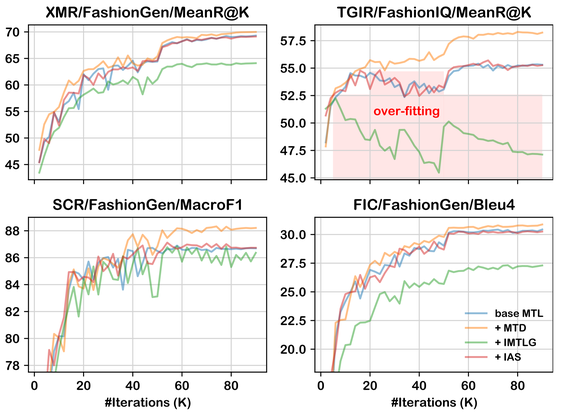
(a) Contrastive Mode (b) Fusion Mode (c) Generative Mode

Quantitative Results



Further Analysis

Methods	#Params (%)	$\mathcal{T}_1$: XMR μ	XMR Δ	$\mathcal{T}_2$: TGIR μ	TGIR Δ	$\mathcal{T}_3$: SCR μ	SCR Δ	$\mathcal{T}_4$: FIC μ	FIC Δ	$\bar{\mu}$	$\bar{\Delta}$
STL	**0.0**	66.30	0.0	51.87	0.0	**90.34**	**0.0**	-	-	52.13	0.0
STL + TSA	+1.35	**69.99**	**+5.56**	52.59	+1.39	90.10	-0.27	-	-	53.25	+1.67
STL + XAA	+14.70	66.30	0.0	53.83	+3.78	89.89	-0.50	**63.70**	**0.0**	68.43	+0.82
STL + TSA + XAA (FAME-ViL(δT))	+15.96	**69.99**	**+5.56**	**55.47**	**+6.94**	90.27	-0.07	63.67	-0.05	**69.85**	**+3.10**
MTL	**-70.43**	57.65	-13.05	49.57	-4.43	85.95	-4.86	-	-	48.29	-5.59
MTL + TSA	-70.11	67.97	+2.52	52.04	+0.33	90.32	-0.02	-	-	52.58	+0.71
MTL + XAA	-67.65	65.87	-0.65	52.59	+1.39	**90.93**	**+0.65**	60.99	-4.25	67.60	-0.72
MTL + TSA + XAA (base MTL)	-67.33	**69.31**	**+4.54**	**55.41**	**+6.82**	90.84	+0.55	**65.17**	**+2.31**	**70.18**	**+3.56**
base MTL + MTD (FAME-ViL)	-67.33	70.00	+5.56	**58.29**	**+12.38**	**91.44**	**+1.22**	65.50	+2.83	**71.31**	**+5.50**
base MTL + MTD + Uniform	-67.33	67.70	+2.11	57.31	+10.49	91.36	+1.13	65.12	+2.23	70.37	+3.99
base MTL + MTD + Round-robin	-67.33	67.79	+2.25	57.47	+10.80	91.35	+1.12	64.87	+1.84	70.37	+4.00
base MTL + IAS [32]	-67.33	69.13	+4.27	55.26	+6.54	90.51	+0.19	63.67	-0.05	69.64	+2.74
base MTL + MTD + IAS [32]	-67.33	**70.11**	**+5.75**	57.97	+11.76	90.88	+0.60	**65.66**	**+3.08**	71.16	+5.30
base MTL + IMTLG [46]	-67.33	64.11	-3.30	47.12	-9.16	90.21	-0.14	55.61	-12.70	64.26	-6.33
base MTL + MTD + IMTLG [46]	-67.33	67.14	+1.27	57.22	+10.31	90.09	-0.28	58.14	-9.56	68.15	+0.44
FAME-ViL (bottleneck dim. = 128)	**-65.14**	70.73	+6.68	58.03	+11.88	**91.54**	**+1.33**	66.20	+3.92	71.63	+5.95
FAME-ViL (bottleneck dim. = 256)	-62.67	71.77	+8.25	58.45	+12.69	91.10	+0.84	66.81	+4.88	72.03	+6.67
FAME-ViL (bottleneck dim. = 512)	-57.73	**72.32**	**+9.08**	**58.51**	**+12.80**	90.96	+0.69	**66.92**	**+5.05**	**72.18**	**+6.91**



Qualitative Results

Text query: Satin cap in black. Adjustable snapback fastening. Tonal hardware. Tonal stitching.

Text query: Short sleeve t-shirt in black. Rib-knit crew-neck collar. Logo printed at front in white and black. Tonal stitching.

Text query: Wide-leg woven cotton sarouel-style trousers in dark navy. Partially elasticized waistband. Pleats at front. Two-pocket styling. Unlined.

Modifying text: is a green t-shirt with a light material, is more colorful.

Modifying text: is blue with a collar and some buttons, is blue and shorter sleeved.

Modifying text: is maroon with a ruffled top, is a dark red cowl-neck and long sleeves.

Images	Ground Truth Captions	Generated Captions
	White logo tank top. Relaxed-fit tank top in white. Ribbed scoopneck collar and armscyes. Logo print at black. Tonal logo embroidered at back hem. Tonal stitching.	White logo tank top. Racer-back tank top in white. Scoopneck collar. Logo printed at front in black. Curved hem. Tonal stitching.
	Black python print shirt. Short sleeve shirt in tones of grey and black. Detailed python scale print throughout with ombre effect at bottom portions. Spread collar. Button closure at front. Tonal stitching. Single-button barrel cuffs with buttoned sleeve placket.	Black paint splatter shirt. Long sleeve shirt in black. Graphic print throughout in white. Spread collar. Button closure at front. Tonal stitching. Single-button barrel cuffs with buttoned sleeve placket.