



Sentiment Analysis of Fashion Related Posts in Social Media

Yifei Yuan

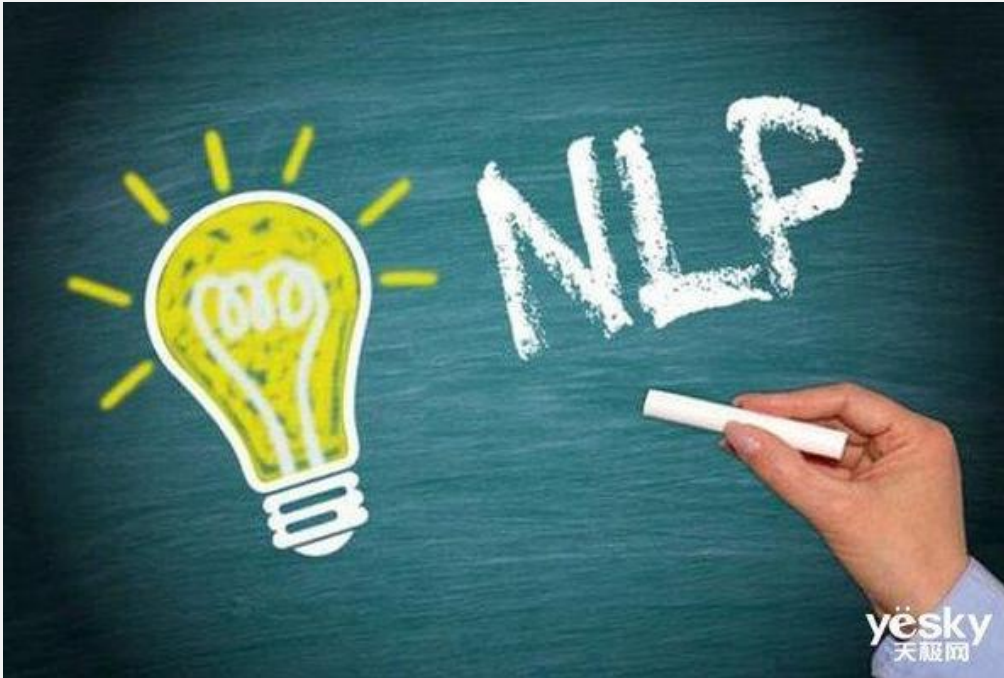
The Chinese University of Hong Kong
Hong Kong SAR
yfyuan@se.cuhk.edu.hk

Wai Lam

The Chinese University of Hong Kong
Hong Kong SAR
wlam@se.cuhk.edu.hk

(WSDM-2022)

Reported by Jia Wang



- 1. Introduction**
- 2. Approach**
- 3. Experiments**

Introduction



Text: I do collaborations (tfcd) with makeups. Send me MP. Would you like to have photos like this? Looking for a personal session? Ask me for information.

Sentiment: Neutral



Text: America the outfit, had to pretend I was happy in this shirt
#worstshirtever #Walmart

Sentiment: Negative



Text: My friend Grace challenged me to share my "best photo" and if this isn't it, I don't know what is. My mom made me wear this.....

Sentiment: Negative



Text: It's a blue feeling
#greece #griechenland
#ocean #oceaneyes #summerootd

Sentiment: Positive

Figure 1: Some examples of our Instagram fashion sentiment analysis dataset. The red box is the fashion items box and the yellow one is the face detection box.



Introduction

In conclusion, the main contributions of this work are as follows:

- Our model detects the user **sentiment polarity** from fashion related posts in social media. A novel framework has been developed, which jointly exploits information from **images, post texts, and fashion attributes**.
- The mutual relationship between the fashion attributes extracted from post images and the post texts is captured via a **mutual attention mechanism**.
- We collect a large-scale dataset of over 12k fashion related social media posts, each includes an image and the corresponding post texts. Extensive experiments are conducted on two datasets to demonstrate the effectiveness of our model.

Approach

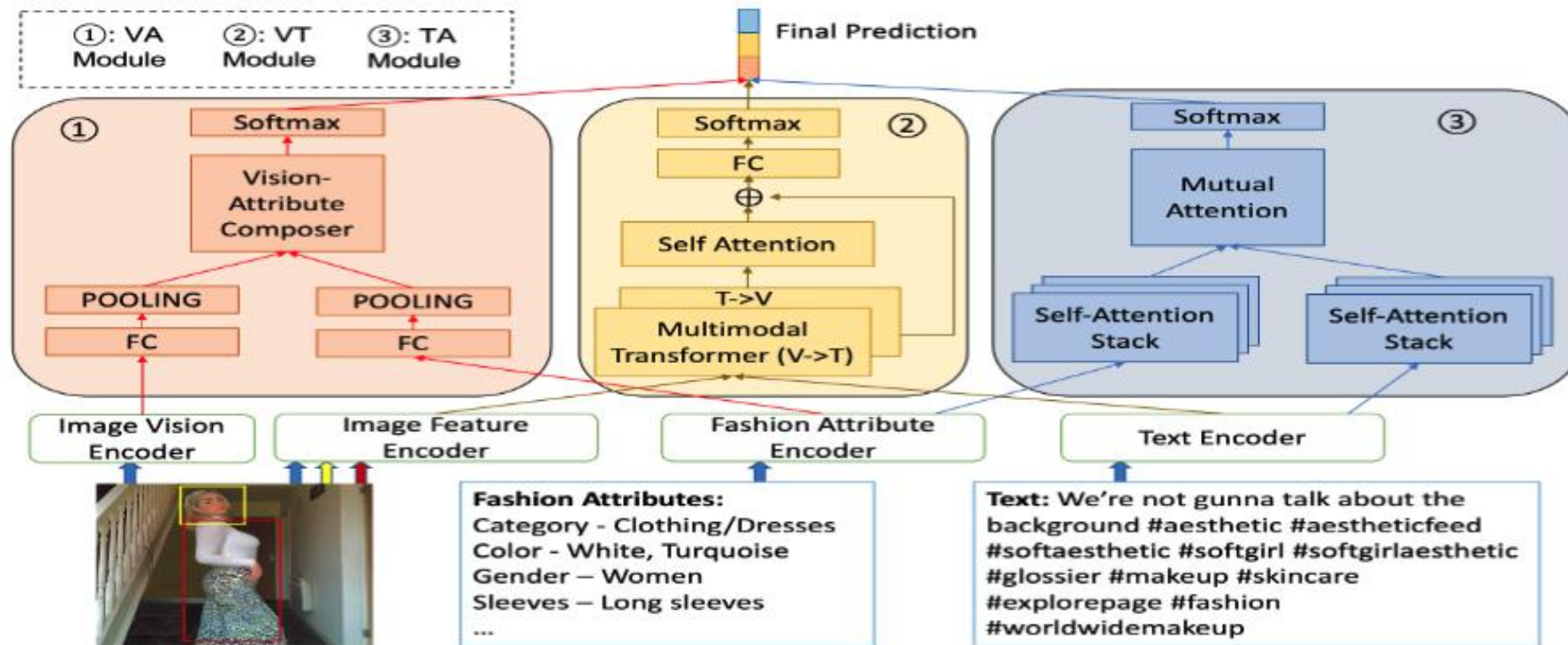


Figure 2: The overall structure of our framework. It consists of a fashion-aware vision composition module (VA module), a fashion-aware text composition module (TA module), and a vision text composition module (VT module). The input is an image annotated with face and fashion item boxes, the corresponding post texts, and the extracted fashion attributes.

Approach

Preprocessing and Basic Encoding

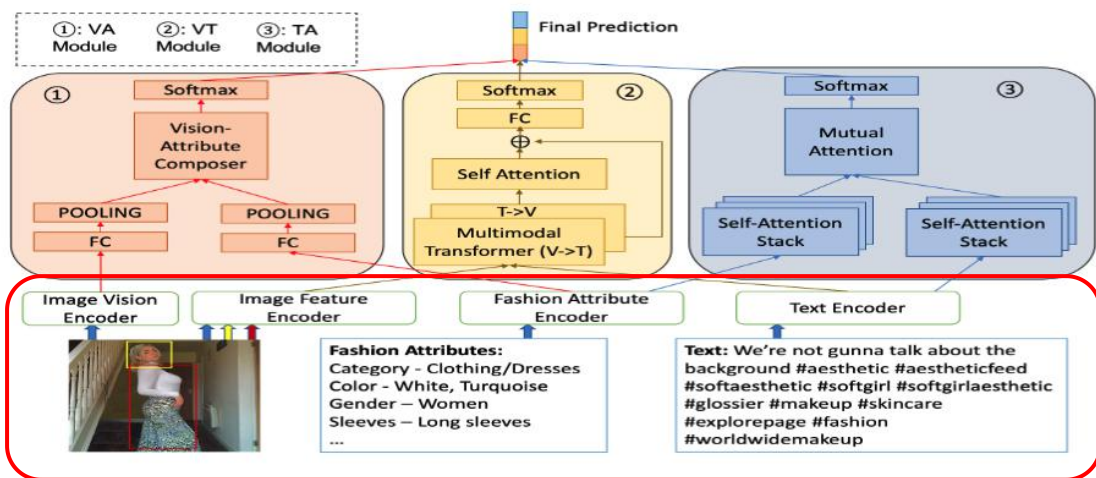


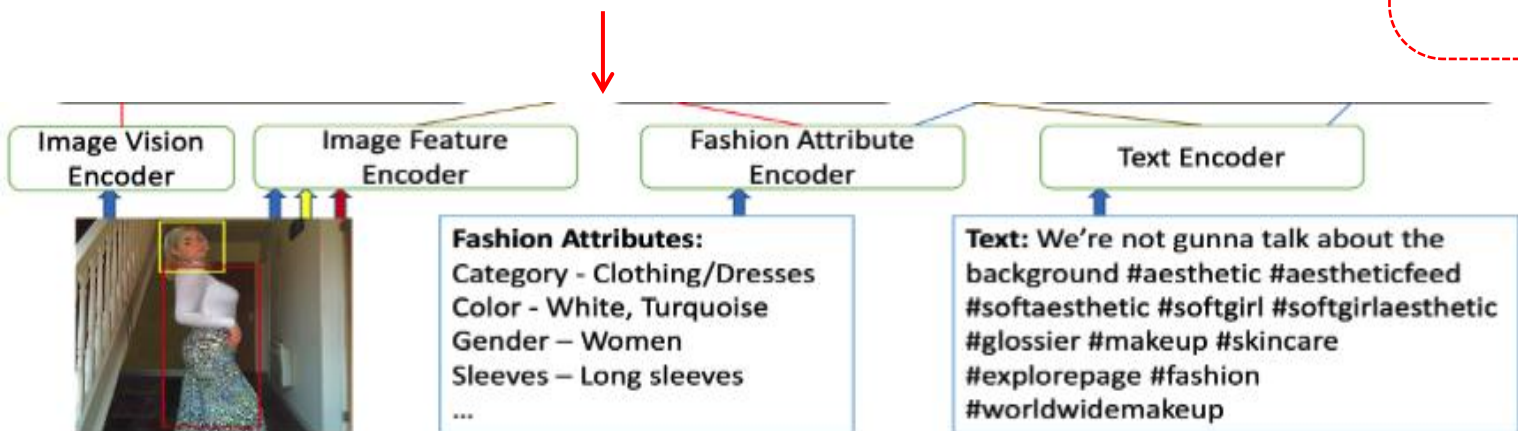
Image Vision Encoder(ResNet)

$$v_i = \text{ImgEnc}(p_i) \in \mathbb{R}^{d_p}$$

Image Feature Encoder
MTCNN (Multi-task Cascaded Convolutional Networks)
YOLOv3, ResNet

$$v'_i = \text{ImgFtEnc}(p_i) \in \mathbb{R}^{l_v \times d_p}$$

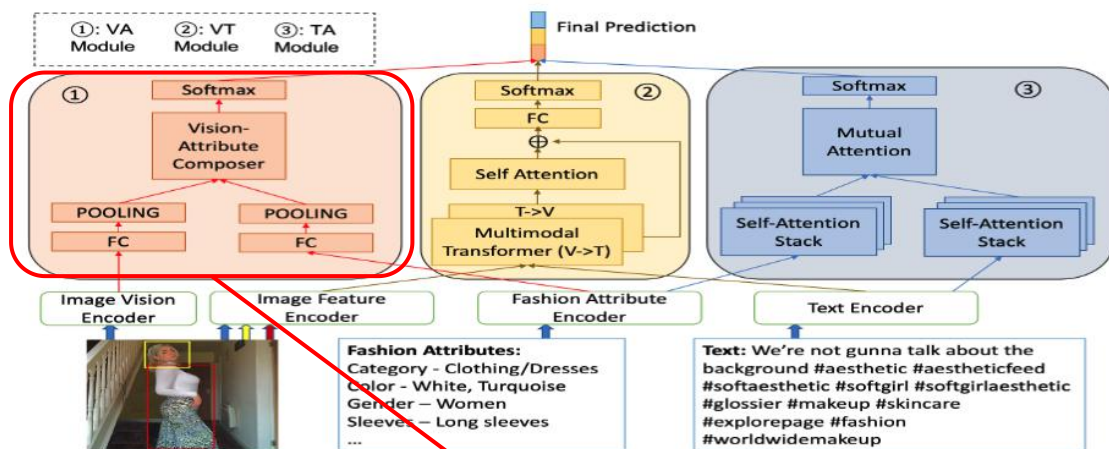
Text Encoder (Glove)

$$e_i = \text{TxtEnc}(t_i) \in \mathbb{R}^{l_t \times d_t}$$


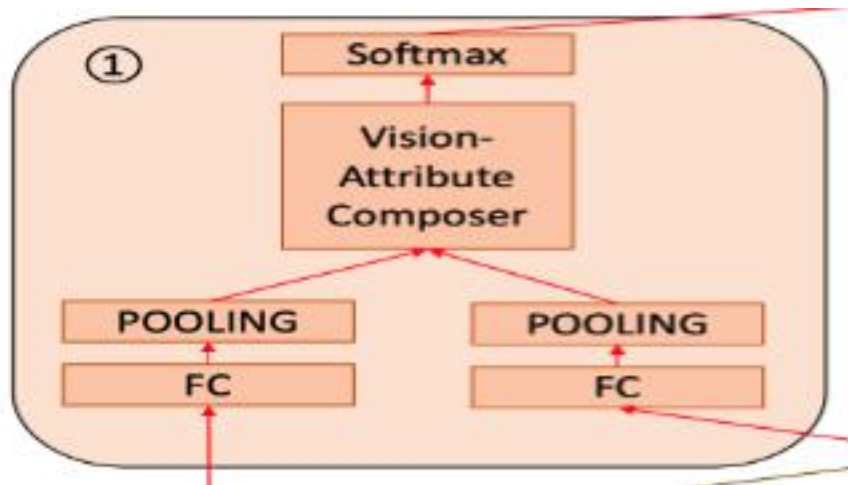
Approach

Fashion-Aware Vision Composition Module

(使用Ximilar4提供的时尚标签工具从图片中提取时尚属性)



TOP CATEGORY	FEATURES	Percentage
Clothing	Age — adult	99.48%
	Color — beige	93.10%
	Design — patterned	90.74%
	Gender — unisex	78.33%
Jackets and Coats	Hood — hood	99.74%
	Length — middle	96.49%
	Material — synthetic	99.72%
	Pattern — stripe	99.51%
	Style — casual	99.06%
	Subcategory — puffer jackets	96.53%
	Subcategory — winter jackets	96.53%



$$f_i = \text{AttrEnc}(a_i) \in \mathbb{R}^{l_a \times d_t}$$

$$c_i = \text{ComposeAE}(\text{FC}(\text{POOLING}(v_i)), \text{FC}(\text{POOLING}(f_i))) \quad (1)$$

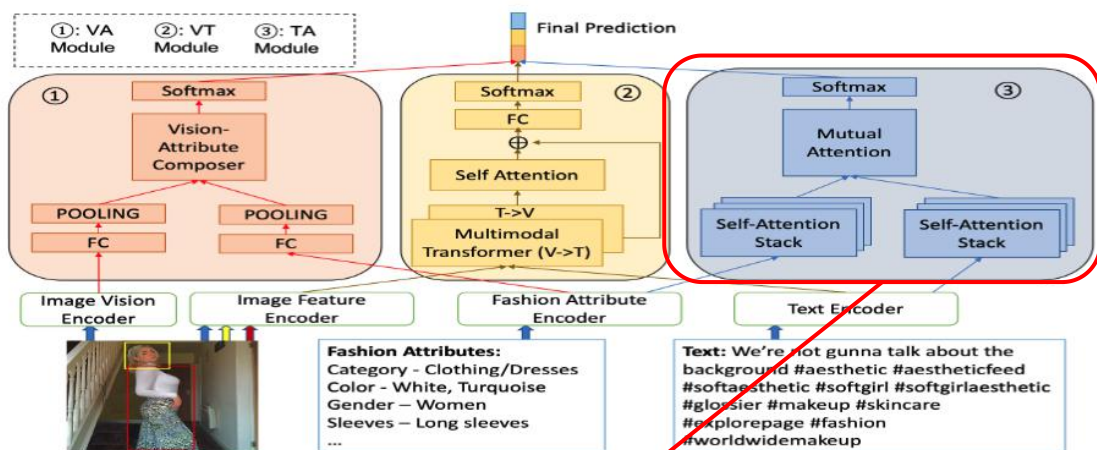
where the composed representation $c_i \in \mathbb{R}^d$.

$$V_{final}^{[1]} = \text{softmax}(c_i) \in \mathbb{R}^{d_c} \quad (2)$$

where d_c is the number of label categories.

Approach

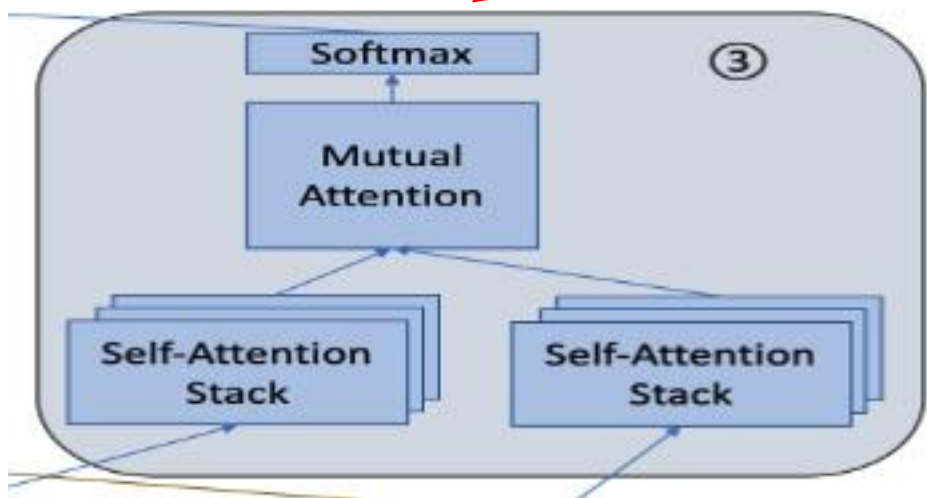
Fashion-Aware Text Composition Module



$$x_i = SAtt(e_i) = Avg(MHAtt(Q, K, V = e_{i,j})_{j=1}^D) \in \mathbb{R}^{l_t \times d_t} \quad (3)$$

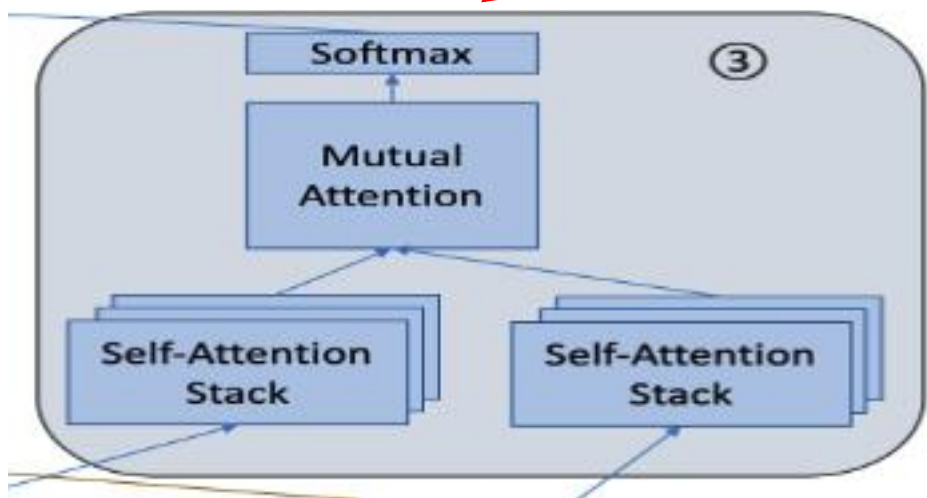
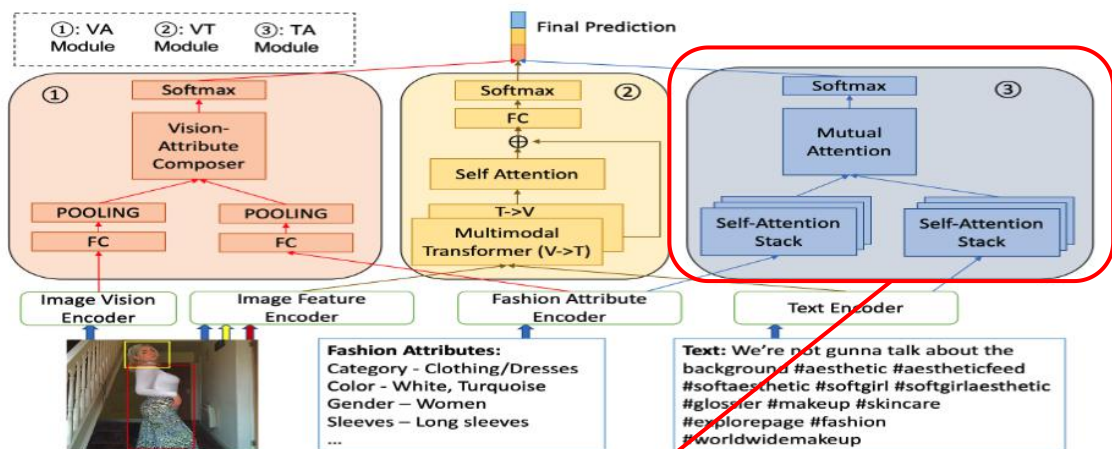
$$y_i = SAtt(f_i) = Avg(MHAtt(Q, K, V = f_{i,j})_{j=1}^D) \in \mathbb{R}^{l_a \times d_t} \quad (4)$$

where D is the number of attention blocks in the self attention stack. $e_{i,j}$ is the text encoding output of the j -th attention block, which also serves as the input of the $j+1$ -th block. $MHAtt$ denotes the multi-head attention mechanism. Note that the initial vector $e_{i,0}$ equals to the text embedding e_i , and $f_{i,0}$ equals to the attribute embedding result f_i .



Approach

Fashion-Aware Text Composition Module



Attribute-to-Text Attention.

$$\alpha_{mn} = \text{softmax}(g(W_1^T [x_{im}; y_{in}] + b_1))_n \quad (5)$$

where α_{mn} is the attention weight between the m -th text feature and the n -th fashion attribute feature. $g(\cdot)$ is a non-linear activation function. $\text{softmax}(\cdot)_n$ denotes the softmax function is performed along n dimensions.

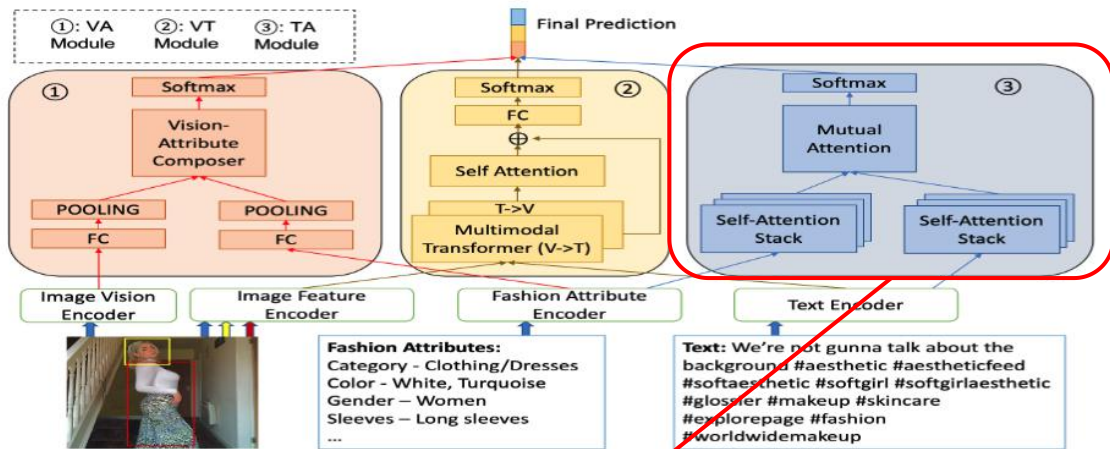
$$x_i^j = \sum_{k=1}^{l_t} \alpha_{jk} x_{ik} \quad (6)$$

$$S(x, a_j) = h(x_i^j, y_{ij}) \quad (7)$$

where $h(\cdot)$ denotes the inner product, y_{ij} is the j -th fashion attribute in the vector.

Approach

Fashion-Aware Text Composition Module

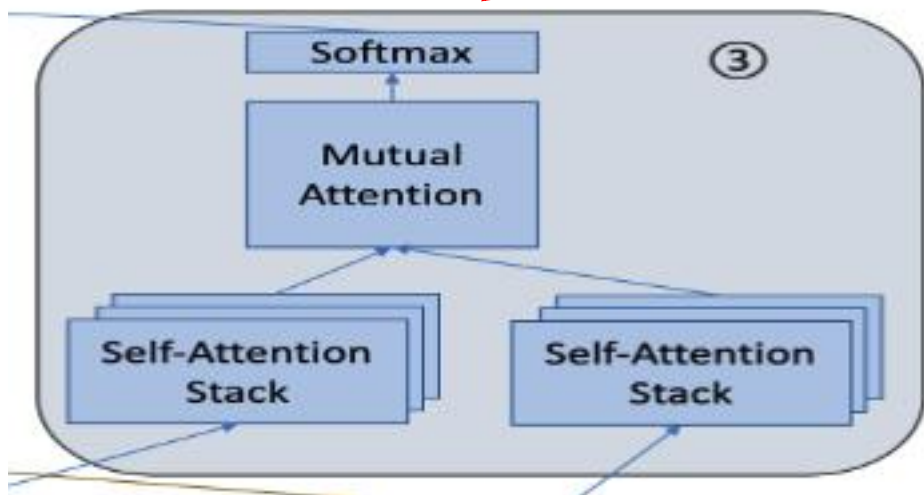


Text-to-Attribute Attention

$$\beta a_m = \text{softmax}(g(W_2^T [\bar{x}_i; y_{im}] + b_2))_m \quad (8)$$

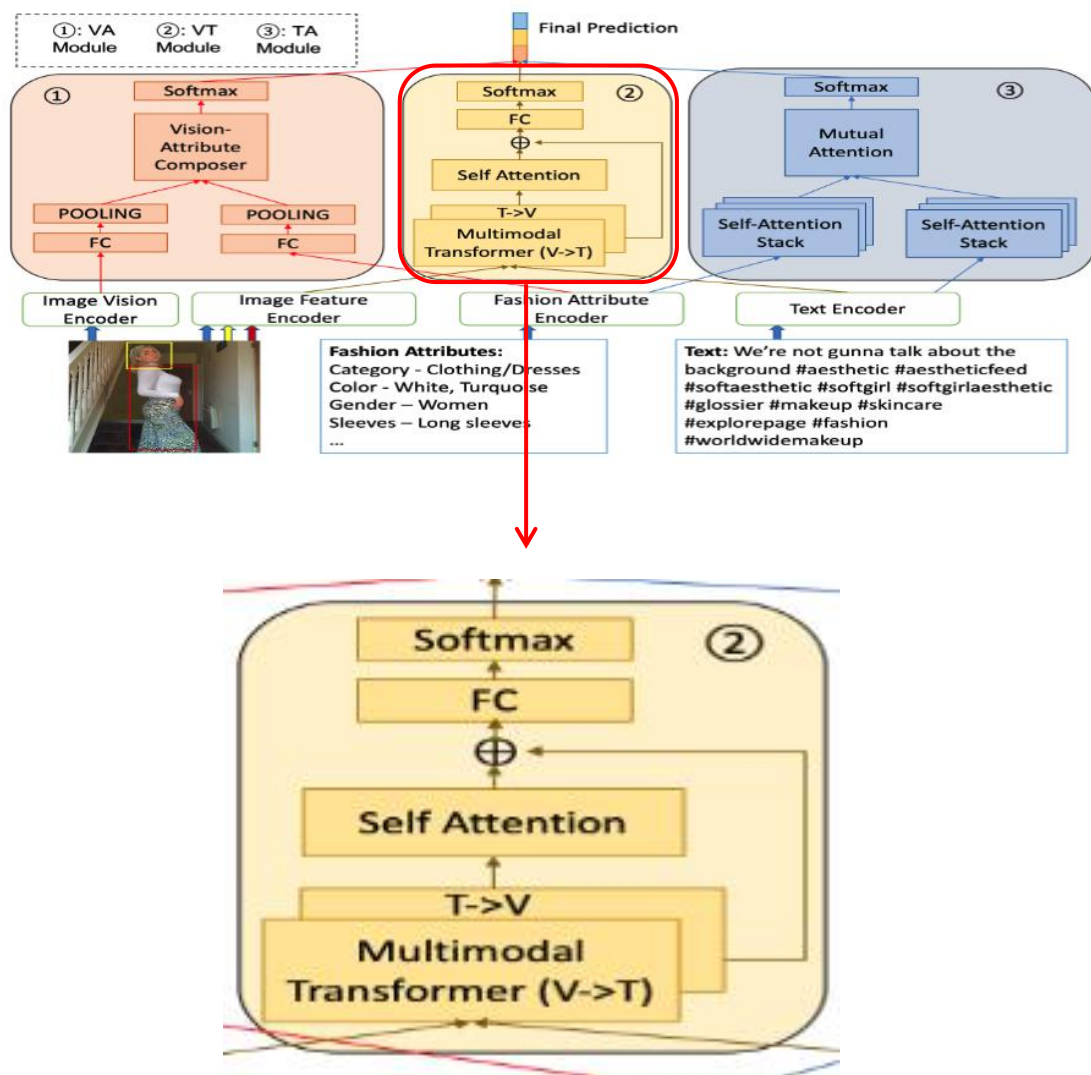
where βa_m is the weight between the m -th fashion attribute and the text representation. \bar{x}_i is calculated by averagely pooling all the feature vectors in x_i . The final partial prediction vector is the weighted sum of the correlation vectors obtained in the last step followed by a fully connected layer with a softmax function.

$$Y_{final}^{[2]} = \text{softmax}(FC(\sum_{a_k \in \{a_e, a_s, \dots\}} \beta a_k S(x, a_k))) \quad (9)$$



Approach

Vision Text Composition Module



$$X_{t \rightarrow v} = MHAtt(Q = W_Q v'_i, K = W_K e_i, V = W_V e_i) \quad (10)$$

$$\hat{Y}_{t \rightarrow v}^{[i]} = MHAtt(LN(Y_{t \rightarrow v}^{[i-1]}), LN(Y_t^{[0]})) + LN(Y_{t \rightarrow v}^{[i-1]}) \quad (11)$$

$$Y_{t \rightarrow v}^{[i]} = f_{\theta}(LN(\hat{Y}_{t \rightarrow v}^{[i]})) + LN(\hat{Y}_{t \rightarrow v}^{[i]}) \quad (12)$$

where LN is the layer normalization function, and $Y_t^{[0]}$ is the text feature. $Y_{t \rightarrow v}^{[0]}$ is initialized as $X_{t \rightarrow v}$. f_{θ} is a positionwise feed-forward sublayer parametrized by θ .

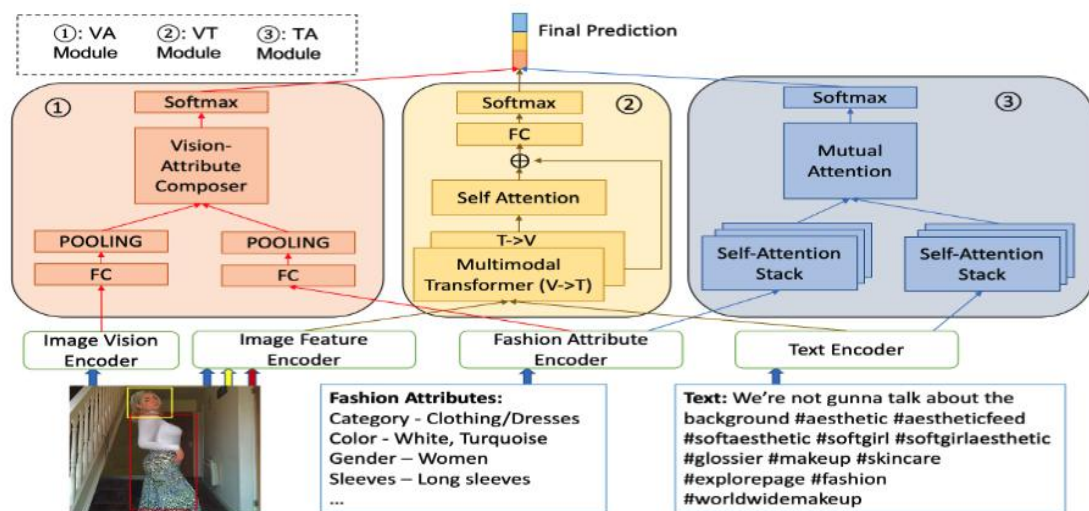
$$Y_{t,v} = [Y_{t \rightarrow v}^{[L]}, Y_{v \rightarrow t}^{[L]}] \quad (13)$$

$$Y_{final}^{[3]} = softmax(FC(SAtt(FC(Y_{t,v})) + Y_{t,v})) \quad (14)$$

where $SAtt$ denotes the self-attention stack. FC is the fully connected layer. L is the number of attention blocks in the stack.

Approach

Partial Prediction Score Combination and Training



$$Y_{final} = w_1 Y_{final}^{[1]} + w_2 Y_{final}^{[2]} + w_3 Y_{final}^{[3]} \quad (15)$$

where w_1, w_2, w_3 are the weights we would like to optimize. During training, we use the softmax cross entropy as the loss function.



Experiments

hashtag	Positive	Negative	Neutral	Total
#fashion	2489	739	1608	4836
#instagood	977	22	413	1412
#ootd	381	120	180	681
#tbt	187	89	111	387
#stupidshirt	59	143	121	323
Total	4387	2727	4947	12061

Table 1: Detailed information of the top 5 hashtags in our dataset.

Experiments

Type	Model	TWMMS1				Our Dataset			
		Acc	Precision	Recall	F1	Acc	Precision	Recall	F1
Unimodal	Text-Only	28.93	17.78	17.50	16.15	57.96±0.66	59.92±1.43	57.28±1.20	56.80±1.06
	Image-Only	15.25	8.55	7.87	8.62	47.04±1.11	46.84±0.76	46.38±0.50	44.58±0.21
	Attribute-Only	6.42	1.12	2.02	1.21	43.95±1.66	39.75±0.44	38.49±0.64	34.94±0.58
Multimodal (2 Modalities)	Early Fusion	31.31	22.66	21.53	19.18	59.68±1.26	62.41±1.41	59.96±1.70	58.91±1.66
	Late Fusion	32.85	23.48	20.61	19.98	60.30±1.06	63.09±2.39	60.76±1.56	60.02±1.63
	TIRG	29.33	18.08	18.16	18.98	62.60±1.49	64.21±1.06	64.95±1.12	63.07±1.16
	ComposeAE	34.45	18.97	19.67	19.99	62.01±0.51	64.31±0.87	62.60±0.85	61.48±0.71
	ViLBERT	36.77	25.68	24.93	22.12	66.25±0.63	66.06±0.56	66.51±0.51	66.23±0.19
	ViLBERT CC	35.78	25.70	24.85	22.26	66.38±0.99	66.13±0.21	67.73±0.20	66.47±0.41
	A2T	11.63	5.87	6.33	5.78	63.18±0.83	66.16±1.62	63.27±1.22	62.55±1.07
	T2A	8.36	2.86	3.42	2.90	62.94±0.83	64.61±0.20	64.89±0.37	63.17±0.40
Multimodal (3 Modalities)	M ³ H-Att	29.27	17.16	17.88	16.16	58.08±2.21	59.85±2.51	58.36±2.41	56.86±2.29
	Ours	37.60	26.82	26.13	24.20	67.58±0.83	68.09±0.76	67.13±0.76	67.56±0.77

Table 2: Experimental results on two datasets



Experiments

Model	Acc	Precision	Recall	F1
VA module	48.26	47.94	48.18	46.71
TA module	63.93	65.72	63.69	62.83
VT module	66.58	66.41	67.96	66.67
w/o VA	67.09	67.94	67.15	66.70
w/o TA	66.07	66.45	65.92	66.18
w/o VT	64.93	66.68	62.98	64.28

Table 3: Ablation study of our proposed model

Experiments

 <p>Text: trying to look like a vampire but the stylist seems failed#modelinmilan #uglyfashion #stylist #collagear Attributes: Sleeves--sleeveless, Material--mesh/transparent, Color--black, Gender: women, Age--adult, Neckline--low cut, Style--elegant Label: Negative</p>	 <p>Text: #sexiestme #fashiondisaster #krishnatomar08 #punjabibrand #dumlight #vulger #mtell #brownboys #anchor #youtuber #youtubeindia Attributes: Category--Clothing/Upper, Style--casual, Cut--fastening, Subcategory-- shirts, Color--grey, Gender--men, Age--adult, Material--corduroy Label: Positive</p>	 <p>Text: getting my daily dose of sea minerals , a bit wierd, isn't it?#DirtyHair #Pose #Photoshooting #LifestyleBlogger Attributes: Category--Clothing/Jackets and Coats, Layers--1st layer, Style--elegant Top Category--Clothing Label: Negative</p>	 <p>Text: Like a barbie doll Attributes: Category--Clothing/Upper, Layers--1st layer, Style--casual, Color--dark blue and navy, Gender--men, Age--adult, Sleeves--long sleeves, Pattern--plain Label: Positive</p>
<p>T-O: Neg TIRG: Neg VA Module: Neu Ours: Neg I-O: Pos ViLBERT: Neu TA Module: Neu A-O: Pos M3H-Att: Neg VT Module: Neg</p>	<p>T-O: Neg TIRG: Neu VA Module: Neg Ours: Pos I-O: Pos ViLBERT: Neu TA Module: Neu A-O: Neu M3H-Att: Neu VT Module: Neu</p>	<p>T-O: Neu TIRG: Neg VA Module: Pos Ours: Neg I-O: Pos ViLBERT: Neg TA Module: Neg A-O: Pos M3H-Att: Neg VT Module: Neg</p>	<p>T-O: Neu TIRG: Pos VA Module: Pos Ours: Pos I-O: Neu ViLBERT: Neg TA Module: Pos A-O: Neu M3H-Att: Neu VT Module: Neu</p>

Figure 3: Case study of different methods

Experiments

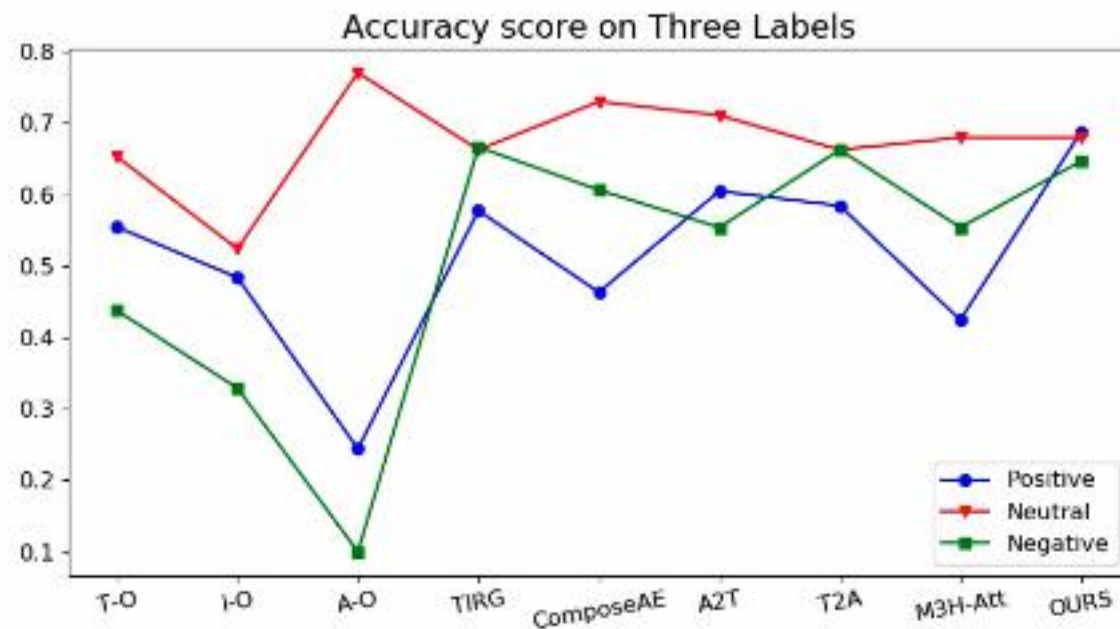


Figure 4: Accuracy score of different methods on three sentiment categories. T-O, A-O, I-O denote the three unimodal methods respectively.



Thanks !