

---

# Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis

Yan Ling, Jianfei Yu\* , and Rui Xia\*

School of Computer Science and Engineering, Nanjing University of Science and Technology, China

ACL 2022 (long paper)

---

Multimodal Aspect-Based Sentiment Analysis task는 Aspect-Based Sentiment Analysis를 확장하여 텍스트와 시각적인 모달리티를 결합하여 특정 측면(Aspect)에 대한 감성(sentiment)을 분석하는 과제

일반적으로, Aspect-Based Sentiment Analysis는 텍스트 문장에 대해 감성을 분석하는 작업

하지만 Multimodal Aspect-Based Sentiment Analysis는 텍스트 외에도 이미지나 비디오와 같은 시각적인 콘텐츠를 함께 분석

Image	
Text	Sergio Ramos chosen as the best player of UCL final
Output	(Sergio Ramos, Positive) (UCL, Neutral)

Table 1: An example of the MABSA task

## **MABSA TASK**

### **Multimodal Aspect Term Extraction (MATE)**

-> text-image쌍이 주어졌을때 본문에 언급된 모든 Aspect Term을 추출

### **Multimodal Aspect-oriented Sentiment Classification (MASC)**

-> 추출된 Term에서 감정을 분류하는것에 목표

### **Joint Multimodal Aspect-Sentiment Analysis (JMASA)**

-> MATE와 MASC를 연관지어 Aspect Term을 추출하고 그 Term에서 감정을 분류

## 한계점

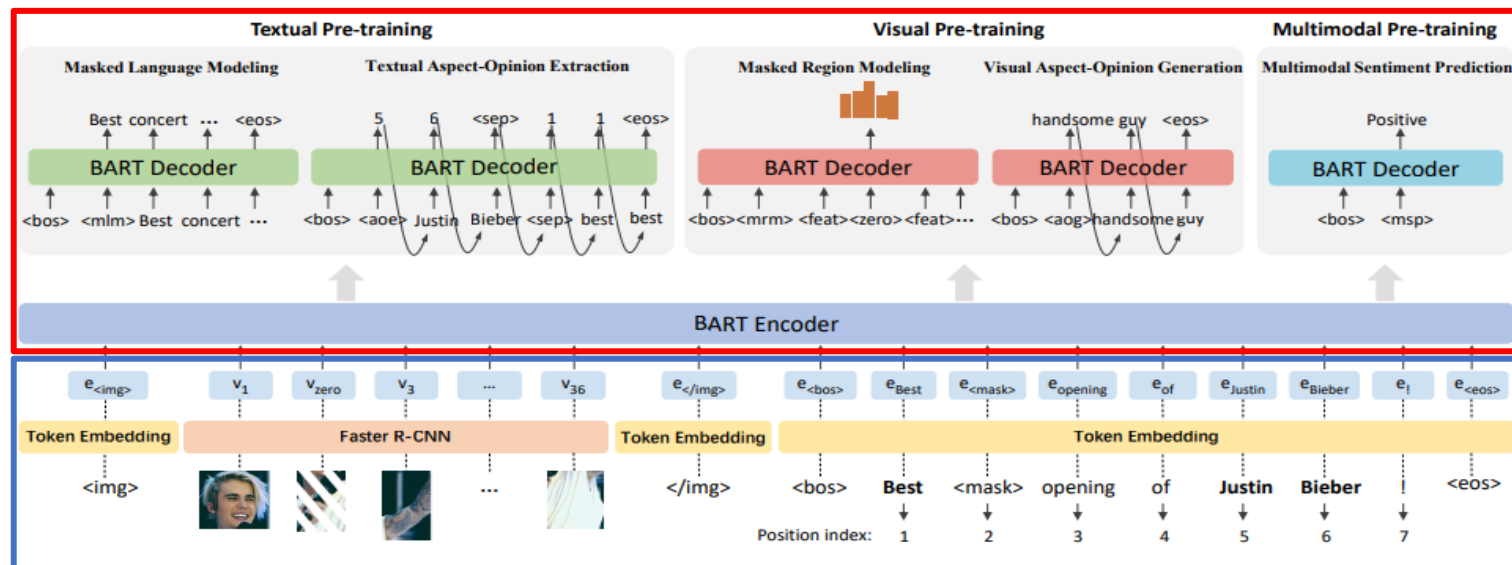
일반적으로 이전에는 MATE, MASC를 연구하는데 유니 모달로 각각 텍스트에는 BERT, 이미지는 ResNet과 을 사용하여 각각 텍스트와 시각적 특성을 추출하는 데 초점을 맞춤

그나마 존재하는 연구도 vision-language understanding tasks에 초점을 맞추고 세부적인 측면에서 이해하는데 한계가 있음

-> 이 문제를 해결하기 위해, 본 논문에서는 MABSA를 위하여 BART기반으로 생성 다중 모달 아키텍처를 구축을 제안

## 전체의 모델 구조도

-> text-image쌍이 주어졌을때 **Feature extractor**와 **Encoder and Decoder** 부분 크게 두 파트로 볼 수 있음



-> Encoder and Decoder

-> Feature Extractor

Figure 1: Overview of our Vision-Language Pre-Training framework for MABSA

## Feature Extractor

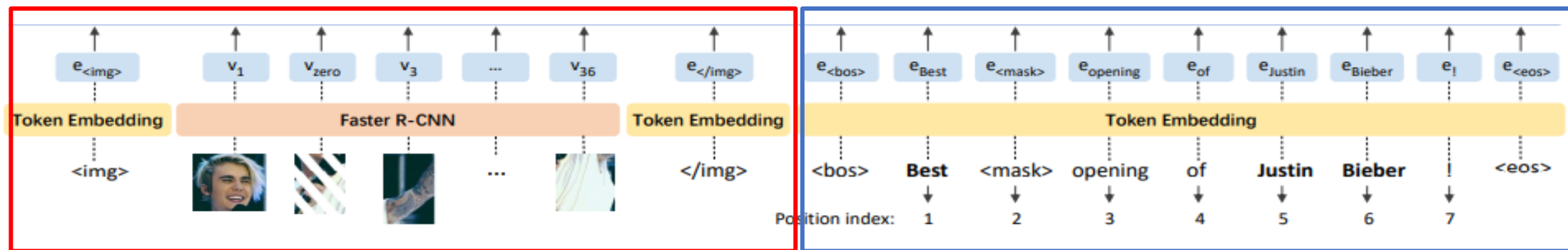
### Image Representation

-> **Faster R-CNN**을 사용하여 높은 신뢰도를 가진 **36개의 영역**을 추출

**Masked Region Modeling task**에 사용하기 위해 **Semantic class distribution**을 유지

### Text Representation

-> **Token Embedding**을 통해 입력 받은 문장을 **Embedding** 형태로 변경



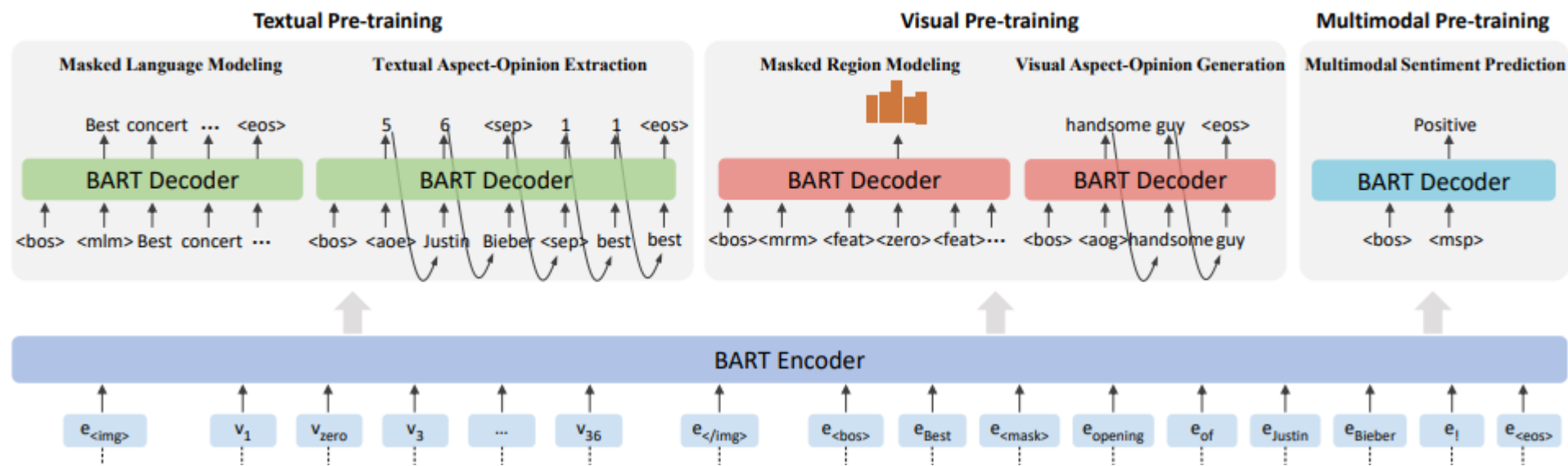
## Encoder-Decoder

### Encoder

-> Feature Extraction에서 나온 Image 부분은  $\langle \text{img} \rangle, \langle / \text{img} \rangle$  로 시작과 끝을 구분 하고 텍스트 부분은  $\langle \text{bos} \rangle, \langle \text{eos} \rangle$ 로 구분

### Decoder

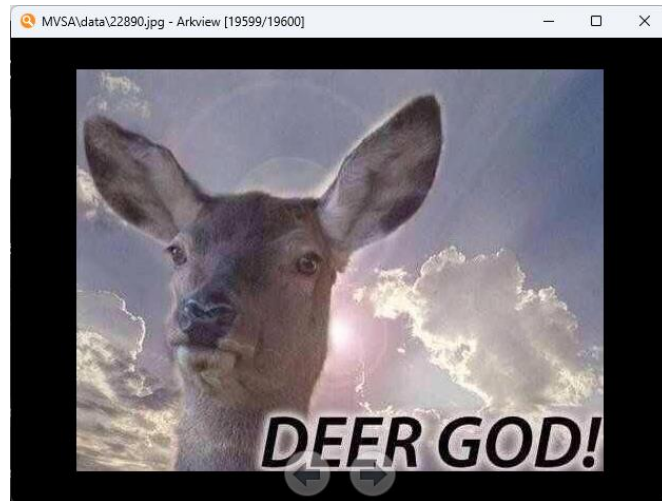
-> 문장 생성의 시작 부분을  $\langle \text{bos} \rangle$ 토큰으로 시작하며 각각 Task마다 뒤에 MLM ( $\langle \text{mlm} \rangle$ ), AOE( $\langle \text{aoe} \rangle$ ), MRM( $\langle \text{mrm} \rangle$ ),AOG( $\langle \text{aog} \rangle$ )로 구분



## Dataset

### MVSA-Multi Dataset

Img :



Text: im dead

Label: negative,negative[text]

Sentiment	#Image-Text Pairs	#Aspects	#Opinions	#Words
Positive	11903	10593	22752	215044
Neutral	4107	3756	7567	74456
Negative	1500	1016	2956	25211

Table 2: The statistics of the MVSA-Multi Dataset. #Apects and #Opinions are the number of aspect terms and opinion terms we extract from the dataset by the rule-based methods introduced in Section 3.3.1.



## Textual Pre-training

### Masked Language Modeling (MLM)

-> **BERT**에서 사용한 방식을 똑같이 채용 하여 입력 텍스트 토큰을 **15%**의 확률로 무작위로 마스킹

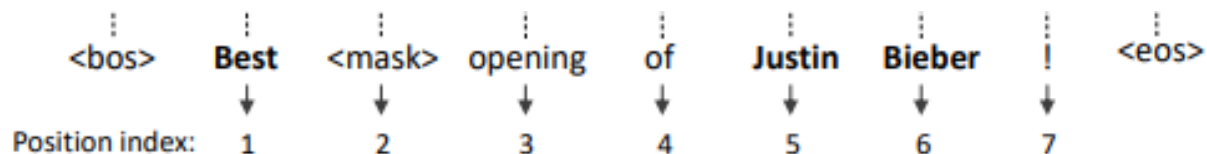
-> **BERT**에서 사용한 **Loss function**을 사용

$$\mathcal{L}_{MLM} = -\mathbb{E}_{X \sim D} \sum_{i=1}^T \log P(e_i | e_{<i}, \tilde{X})$$

## Textual Pre-training

### Textual Aspect-Opinion Extraction

**AOE**작업을 인덱스 생성 작업으로 정의 하여  $Y = [a_1^s, a_1^e, \dots, a_m^s, a_m^e, < sep >, o_1^s, o_1^e, \dots, o_N^s, o_N^e, < eos >]$ 를 생성하는 것을 목표



예시)  $Y = [5, 6<sep>1, 1]$  속성어인 **Justin Bieber**의 시작과 끝 index인 5, 6과 그에관한 의견인 **Best**의 시작과 끝 index인 1, 1

$$\begin{aligned} \mathbf{h}_t^d &= \text{Decoder}(\mathbf{H}^e; Y_{<t}), \\ \bar{\mathbf{H}}_T^e &= (\mathbf{W} + \mathbf{H}_T^e)/2, \\ P(y_t) &= \text{Softmax}([\bar{\mathbf{H}}_T^e; \mathbf{C}^d]\mathbf{h}_t^d), \end{aligned}$$

$$\mathcal{L}_{AOE} = -\mathbb{E}_{X \sim D} \sum_{t=1}^O \log P(y_t | Y_{<t}, X),$$

## Visual Pre-training

### Masked Region Modeling (MRM)

-> 아래 예시 그림과 같이 입력 이미지 벡터를 **15%**의 확률로 무작위로 마스킹 (**zero vector**)



-> **Decoder**의 입력 부분으로 시작토큰은 **<bos><mrm>** 마스킹된 부분은 **<zero>**, 나머지 부분은 **<feat>**으로 입력하여

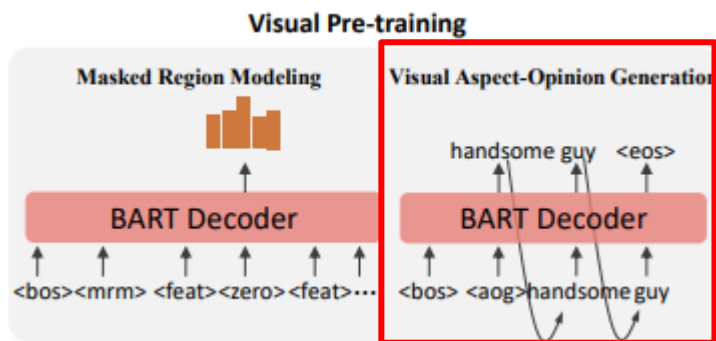
MLP classifier가 zero부분의 semantic class distribution을 예측

$$\mathcal{L}_{MRM} = \mathbb{E}_{X \sim D} \sum_{z=1}^Z D_{KL}(q(v_z) || p(v_z)),$$

## Visual Pre-training

### Visual Aspect-Opinion Generation(AOG)

-> 입력 이미지에서 **aspect-opinion** 쌍을 생성 이때 형용사-명사 쌍(**ANP, Adjective-Noun Pair**) 개념을 사용



ANP에서 handsome guy가 생성되는데 handsome은 aspect , guy는 opinion

AOG task를 Sequence generation Task로 간주하여  $G = \{g_1, \dots, g_{|G|}\}$

$$\mathbf{h}_i^d = \text{Decoder}(\mathbf{H}^e; G_{<i}),$$

$$P(g_i) = \text{Softmax}(\mathbf{E}^T \mathbf{h}_i^d),$$

$$\mathcal{L}_{AOG} = -\mathbb{E}_{X \sim D} \sum_{i=1}^{|G|} \log P(g_i | g_{<i}, X).$$

## Multimodal Pre-training

Multimodal sentiment Prediction(MSP)

-> **MVSA-Multi** 데이터셋을 사용하여 이미지, 텍스트에 대한 감정 분류만 진행

시작토큰으로 **<bos><mSP>**를 사용

$$\mathbf{h}_{msp}^d = \text{Decoder}(\mathbf{H}^e; \mathbf{E}_{msp}),$$

$$P(s) = \text{Softmax}(\text{MLP}(\mathbf{h}_{msp}^d)),$$

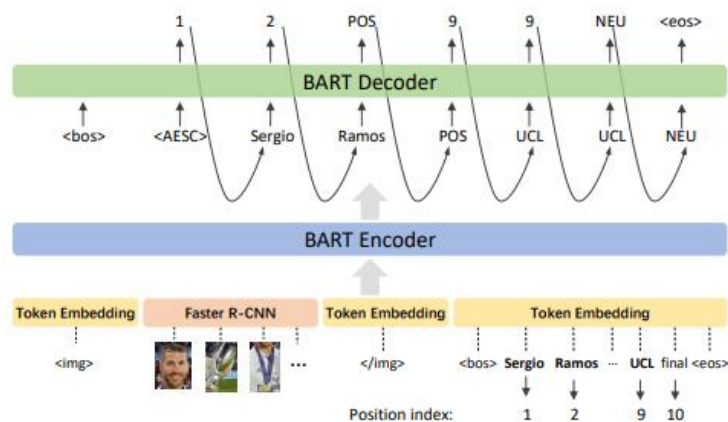
$$\mathcal{L}_{MSP} = -\mathbb{E}_{X \sim D} \log P(s|X),$$

## Full training loss

최종적으로 앞에서 학습한 **Loss**를 전부 더함

$$\mathcal{L} = \lambda_1 \mathcal{L}_{MLM} + \lambda_2 \mathcal{L}_{AOE} + \lambda_3 \mathcal{L}_{MRM} + \lambda_4 \mathcal{L}_{AOG} + \lambda_5 \mathcal{L}_{MSP}$$

## Downstream Tasks



**JMASA**의 경우 예시와 같이 [1, 2,<pos> 9, 9,<neu>,<eos>]. 생성

**Sergio Ramos → Positive UCL → neutral**

Figure 2: An example of downstream task JMASA.  $\langle AESC \rangle$  informs the current task is JMASA.

## JMASA 결과

테이블에서 볼 수 있듯이, **BART**는 텍스트 기반 방법 중에서 가장 좋은 성능

본 논문에서 제안한 모든 사전 훈련 작업을 포함한 전체 모델인 **VLP-MABSA**이 다른 모델에 비해 높은 성능을 가짐을 확인 할수 있다.

	TWITTER-2015			TWITTER-2017		
	P	R	F1	P	R	F1
Text-based methods						
SPAN*	53.7	53.9	53.8	59.6	61.7	60.6
D-GCN*	58.3	58.8	59.4	64.2	64.1	64.1
<b>BART</b>	<b>62.9</b>	<b>65.0</b>	<b>63.9</b>	<b>65.2</b>	<b>65.6</b>	<b>65.4</b>
Multimodal methods						
UMT+TomBERT*	58.4	61.3	59.8	62.3	62.4	62.4
OSCGA+TomBERT*	61.7	63.4	62.5	63.4	64.0	63.7
OSCGA-collapse*	63.1	63.7	63.2	63.5	63.5	63.5
RpBERT-collapse*	49.3	46.9	48.0	57.0	55.4	56.2
JML*	65.0	63.2	64.1	66.5	65.5	66.0
<b>VLP-MABSA</b>	<b>65.1</b>	<b>68.3</b>	<b>66.6</b>	<b>66.9</b>	<b>69.2</b>	<b>68.0</b>

Table 4: Results of different approaches for JMASA. \* denotes the results are from [Ju et al. \(2021\)](#).

## MATE, MASC 결과

**MATE**의 경우 **TWITTER-2017**을 제외하고 높은 성능이 나옴을 확인 할수있다

Methods	TWITTER-2015			TWITTER-2017		
	P	R	F1	P	R	F1
RAN*	80.5	81.5	81.0	90.7	90.0	90.3
UMT*	77.8	81.7	79.7	86.7	86.8	86.7
OSCGA*	81.7	82.1	81.9	90.2	90.7	90.4
JML-MATE*	83.6	81.2	82.4	<b>92.0</b>	90.7	91.4
VLP-MABSA	<b>83.6</b>	<b>87.9</b>	<b>85.7</b>	90.8	<b>92.6</b>	<b>91.7</b>

Table 5: Results of different approaches for MATE. \* denotes the results are from [Ju et al. \(2021\)](#).

**MASC**의 경우 **TWITTER-2015**을 제외하고 높은 성능이 나옴을 확인 할수있다.

Methods	TWITTER-2015		TWITTER-2017	
	Acc	F1	Acc	F1
TomBERT	77.2	71.8	70.5	68.0
CapTrBERT	78.0	73.2	72.3	70.2
JML-MASC	<b>78.7</b>	-	72.7	-
VLP-MABSA	78.6	<b>73.8</b>	<b>73.8</b>	<b>71.8</b>

Table 6: Results of different approaches for MASC. Note that JML-MASC only evaluates on the aspects correctly predicted by JML-MATE while the other methods evaluate on all the golden aspects.



## 각각 방법을 Pretrain 했을때 결과

더 많은 사전 훈련 작업을 추가할 때 대부분의 지표에 대한 성능이 일반적으로 향상

여기서 **Week supervision**은 랜덤으로 **200**개의 샘플을 추출하여 학습

최종적으로 **MSP**를 추가하여 학습했을때 대부분의 **TASK**에서 성능향상을 보이나

특히 **MASC**의경우 더 많은 향상을 확인 할수있음

		TWITTER-2015			TWITTER-2017		
		JMASA	MATE	MASC	JMASA	MATE	MASC
Full supervision	w/o pre-training	65.31	84.80	76.81	66.10	90.67	72.78
	+ <b>T</b> <sub>MLM</sub>	65.44	84.91	77.08	66.27	91.00	72.82
	+ <b>T</b> <sub>AOE</sub>	65.92	85.43	77.48	67.12	91.75	72.89
	+ <b>V</b> <sub>MRM</sub>	65.94	85.49	77.53	67.15	91.72	73.13
	+ <b>V</b> <sub>AOG</sub>	66.38	<b>85.73</b>	77.82	67.66	<b>91.77</b>	73.32
	+ <b>MM</b> <sub>MSP</sub>	<b>66.64</b>	85.66	<b>78.59</b>	<b>68.05</b>	91.73	<b>73.82</b>
Weak supervision	w/o pre-training	39.79	69.33	57.40	49.12	80.48	61.04
	+ <b>T</b> <sub>MLM</sub>	40.42	69.69	58.00	49.69	81.26	61.15
	+ <b>T</b> <sub>AOE</sub>	46.15	79.13	58.32	52.00	84.60	61.46
	+ <b>V</b> <sub>MRM</sub>	46.64	79.49	58.68	52.18	84.47	61.78
	+ <b>V</b> <sub>AOG</sub>	47.79	<b>80.94</b>	59.32	53.16	<b>85.04</b>	62.51
	+ <b>MM</b> <sub>MSP</sub>	<b>51.71</b>	80.69	<b>62.58</b>	<b>55.38</b>	84.88	<b>64.42</b>

Table 7: The results of pre-training tasks on two benchmarks. We evaluate over three tasks JMASA, MATE, and MASC in terms of *F1*, *F1* and *Acc*, respectively. *T*, *V*, and *MM* denote the Textual, Visual, and Multimodal pre-training, respectively. Each row adds an extra pre-training task to the row above it.

## 다른 모델과 비교


Image				
Text	<p>(a) RT @ PearlJam : Eddie and the Faithfull Pearl Jam fans in Buenos Aires . Photo by @ epozzoni # PISA2013</p> <p>(b) RT @ BBCOne : Dear Madonna , THIS is how you wear a cape . # Poldark # Demelza</p> <p>(c) RT @ TrumpDoral : Congratulations to the the new # MissUniverse , Miss Colombia , Paulina Vega !</p> <p>(d) RT @ myfox8 : Charlotte @ hornets visit # Greensboro for D - League meeting</p>			
GT	(Eddie, POS) (Pearl Jam, POS) (Buenos Aires, NEU)	(Madonna, POS) (Poldark, NEU) (Demelza, NEU)	(Miss Colombia, POS) (Paulina Vega, POS)	(Charlotte, NEU) (Greensboro, NEU) (D – League, NEU)
BART	(Eddie, NEU) × (the Faithfull Pearl Jam, NEU) × (Buenos Aires, NEU) ✓	(Madonna, POS) ✓ - × - ×	(Colombia, POS) × (Paulina Vega, POS) ✓	(Charlotte, NEU) ✓ (Greensboro, NEU) ✓ - ×
MM	(Eddie, NEU) × (the Faithfull Pearl Jam, NEU) × (Buenos Aires, NEU) ✓	(Madonna, NEU) × - × (Demelza, NEU) ✓	(Colombia, NEU) × (Paulina Vega, POS) ✓	(Charlotte, NEU) ✓ (Greensboro, NEU) ✓ - ×
VLP	(Eddie, POS) ✓ (Pearl Jam, POS) ✓ (Buenos Aires, NEU) ✓	(Madonna, POS) ✓ (Poldark, NEU) ✓ (Demelza, NEU) ✓	(Miss Colombia, POS) ✓ (Paulina Vega, POS) ✓	(Charlotte, NEU) ✓ (Greensboro, NEU) ✓ (D – League, NEU) ✓

Table 8: Predictions of different methods on four test samples. NEU, POS, and NEG denote Neutral, Positive, and Negative sentiments, respectively.