# Oscar : Object-Semantics Aligned Pre-training for Vision-Language Tasks

**Name 김윤서**

Email diakys2@navere.com

**Study Group vision**

2024/03/19

TRAIN AND TEST

# Contents

- VLP(Vision-language pre-trained models) background
- OSCAR intro
- OSCAR Pre-training

- input/pre-training objective/pre-training corpus/implementation details

- Experiments

TRAIN AND TEST

# VLP Background

| VL-PTM | ① Text encoder | ① Vision encoder | ② Fusion scheme | ③ Pre-training tasks | Multimodal datasets for pre-training |
|---|---|---|---|---|---|
| ② Fusion Encoder | | | | | |
| VisualBERT [2019] | BERT | Faster R-CNN | Single stream | MLM+ITM | COCO |
| Uniter [2020] | BERT | Faster R-CNN | Single stream | MLM+ITM+WRA+MRFR+MRC | CC+COCO+VG+SBU |
| OSCAR [2020c] | BERT | Faster R-CNN | Single stream | MLM+ITM | CC+COCO+SBU+Flickr30k+VQA |
| InterBert [2020] | BERT | Faster R-CNN | Single stream | MLM+MRC+ITM | CC+COCO+SBU |
| ViLBERT [2019] | BERT | Faster R-CNN | Dual stream | MLM+MRC+ITM | CC |
| LXMERT [2019] | BERT | Faster R-CNN | Dual stream | MLM+ITM+MRC+MRFR+VQA | COCO+VG+VQA |
| VL-BERT [2019] | BERT | Faster R-CNN+ ResNet | Single stream | MLM+MRC | CC |
| Pixel-BERT [2020] | BERT | ResNet | Single stream | MLM+ITM | COCO+VG |
| Unified VLP [2020] | UniLM | Faster R-CNN | Single stream | MLM+seq2seq LM | CC |
| UNIMO [2020b] | BERT, RoBERTa | Faster R-CNN | Single stream | MLM+seq2seq LM+MRC+MRFR+CMCL | COCO+CC+VG+SBU |
| SOHO [2021] | BERT | ResNet + Visual Dictionary | Single stream | MLM+MVM+ITM | COCO+VG |
| VL-T5 [2021] | T5, BART | Faster R-CNN | Single stream | MLM+VQA+ITM+VG+GC | COCO+VG |
| XGPT [2021] | transformer | Faster R-CNN | Single stream | IC+MLM+DAE+MRFR | CC |
| Visual Parsing [2021] | BERT | Faster R-CNN + Swin transformer | Dual stream | MLM+ITM+MFR | COCO+VG |
| ALBEF [2021a] | BERT | ViT | Dual stream | MLM+ITM+CMCL | CC+COCO+VG+SBU |
| SimVLM [2021b] | ViT | ViT | Single stream | PrefixLM | C4+ALIGN |
| WenLan [2021] | RoBERTa | Faster R-CNN + EffcientNet | Dual stream | CMCL | RUC-CAS-WenLan |
| ViLT [2021] | ViT | Linear Projection | Single stream | MLM+ITM | CC+COCO+VG+SBU |
| ② Dual Encoder | | | | | |
| CLIP [2021] | GPT2 | ViT, ResNet | | CMCL | self-collected |
| ALIGN [2021] | BERT | EffcientNet | | CMCL | self-collected |
| DeCLIP [2021b] | GPT2, BERT | ViT, ResNet, RegNetY-64GF | | CMCL+MLM+CL | CC+self-collected |
| ② Fusion Encoder+ Dual Encoder | | | | | |
| VLMo [2021a] | BERT | ViT | Single stream | MLM+ITM+CMCL | CC+COCO+VG+SBU |
| FLAVA [2021] | ViT | ViT | Single stream | MMM+ITM+CMCL | CC+COCO+VG+SBU+RedCaps |

* A Survey of Vision-Language Pre-Trained Models(2022) https://arxiv.org/abs/2202.10936

1) 텍스트, 이미지 각각 인코딩 -> 2) 두 모달의 상호작용을 표현할 수 있는 구조 설계 -> 3) 사전학습 -> 4) task에 맞게 fine-tuned

TRAIN AND TEST

# VLP Background

## ② 두 모달 간 상호작용을 모델링하는 구조 설계

1) Fusion Encoder : input으로 텍스트 임베딩과 이미지 임베딩을 동시에(concat) + special embedding
- single stream : 단일 트랜스포머 인코더 + 자체 attention ex) visualBERT, V-L BERT, OSCAR
- dual stream : cross attention(Q/K,V) ex) ViL-BERT, LXMERT, ALBEF

2) Dual Encoder : 두 개의 단일 인코더 사용
- Attention layer 또는 dot product 같은 직관적인 방법사용 -> 동일 의미 공간 투영 -> 유사성 점수 계산

3) Combination of fusion Encoder & dual Encoder

↓ 이해작업 good

↓ 검색 작업 good, 덜 무거움(미리 임베딩가능하기 때문)

TRAIN AND TEST

# VLP Background

## ③ 사전학습 방법

1) Cross-modal Masked language modeling(MLM)
: BERT와 유사한 텍스트 마스킹 기법. 마스킹되지 않은 텍스트 토큰과 이미지를 기반으로 마스킹 토큰 예측

$$L_{\mathrm{MLM}} = -\mathbb{E}_{(W,V)\in\mathcal{D}} \log P_\theta \left(w_m | w_{\setminus m}, V\right),$$

2) Cross-modal masked Region prediction(MRP)
: 일부 ROI(features)을 0으로 마스킹, 다른 이미지 특징을 기반으로 이를 예측
- Masked Region classification(MRC) : 마스킹 영역의 의미 클래스 예측(분류 task)
- Masked Region Feature Regression(MRFR) : 마스킹 영역특징을 원래 영역 특징으로 회귀시킴

3) Image-text matching(ITM)
: 이미지-텍스트 간 상관관계 학습. 이미지-텍스트가 일치하는지 여부로 학습

$$\begin{aligned}\mathcal{L}_{\mathrm{ITM}} = -\,\mathbb{E}_{(W,V)\in\mathcal{D}} \Big[ & y \log s_\theta \left(\boldsymbol{h}_{w_{[\mathrm{CLS}]}}, \boldsymbol{h}_{v_{[\mathrm{IMG}]}}\right) \\ & + (1-y) \log \left(1 - s_\theta \left(\boldsymbol{h}_{w_{[\mathrm{CLS}]}}, \boldsymbol{h}_{v_{[\mathrm{IMG}]}}\right)\right)\Big]\end{aligned}$$

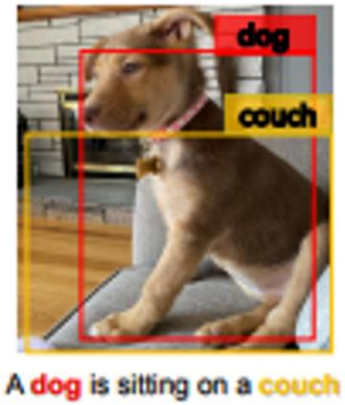
4) Cross-modal(contrastive learning CMCL)
: 일치하는 이미지-텍스트 쌍의 임베딩을 함께 이동, 일치안하는건 멀리 이동시킴.

TRAIN AND TEST

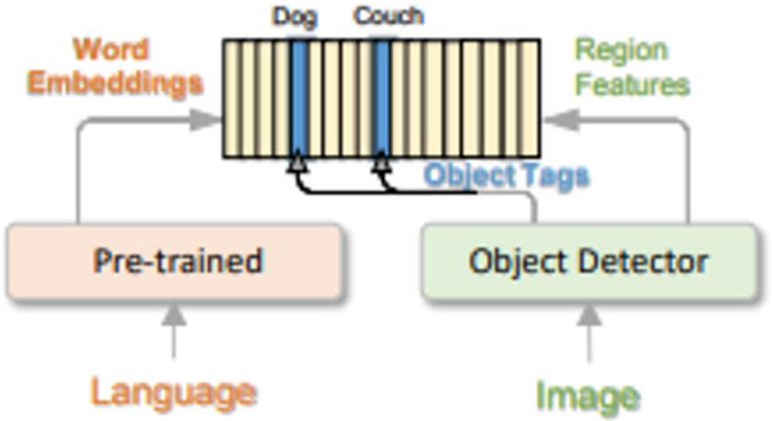
# OSCAR Intro

- OSCAR은 Fusion Encoder (단일 인코더) 계열 모델임.
- OSCAR은 단일 인코더와 단순 attention 메커니즘을 사용하는 것의 문제점을 개선하고자 한다.
<P>
1) 이미지 영역과 텍스트 간의 명시적인 정렬 정보 부재(레이블이 따로 지정X-> weakly-supervised learning problem)
2) 시각적인 영역은 종종 over-sampling 되어, 노이즈가 많고 모호함(이미지가 서로 겹치면 특징애매)

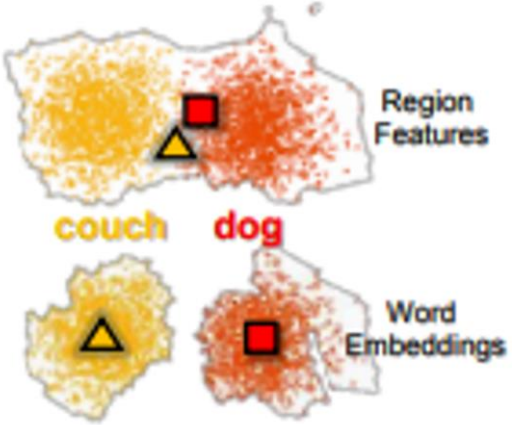
- OSCAR은 이미지에서 감지된 객체 태그를 도입하여 시각과 언어 간의 의미적인 정렬 학습을 용이하게 함
- 훈련 샘플을 1) word sequence 2) a set of object tags 3) set of image region features로 구성된 트리플로 새로 정의 (보통 V-L 데이터셋은 이미지-텍스트 쌍임)



(a) Image-text pair



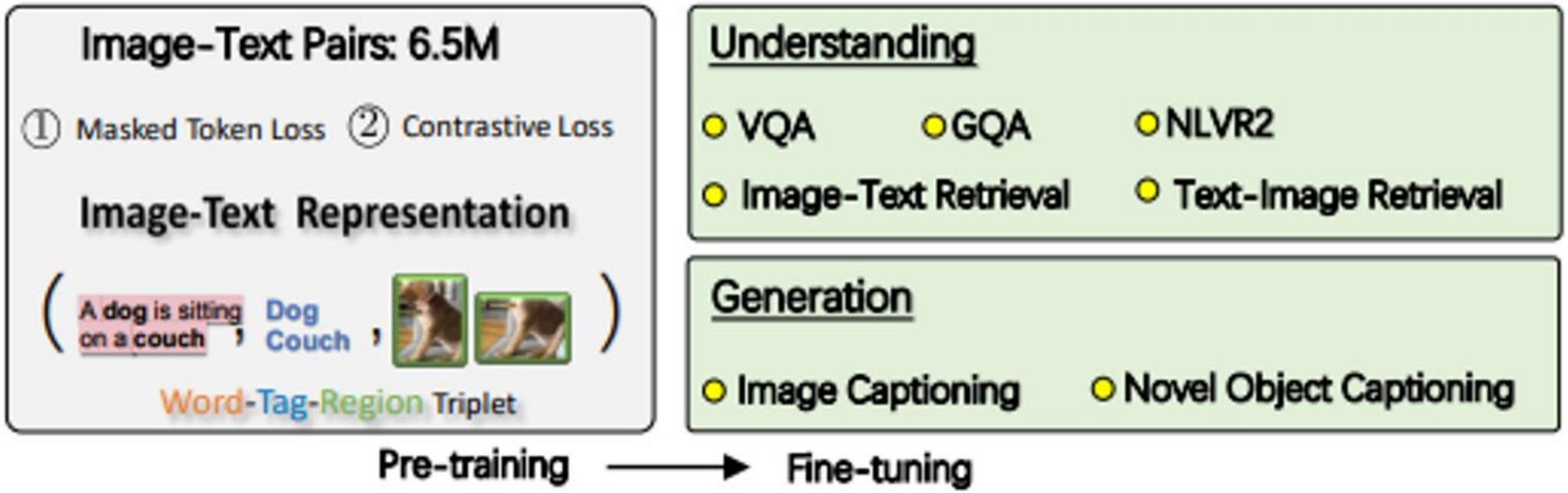
(b) Objects as anchor points



(c) Semantics spaces

TRAIN AND TEST

# OSCAR Intro



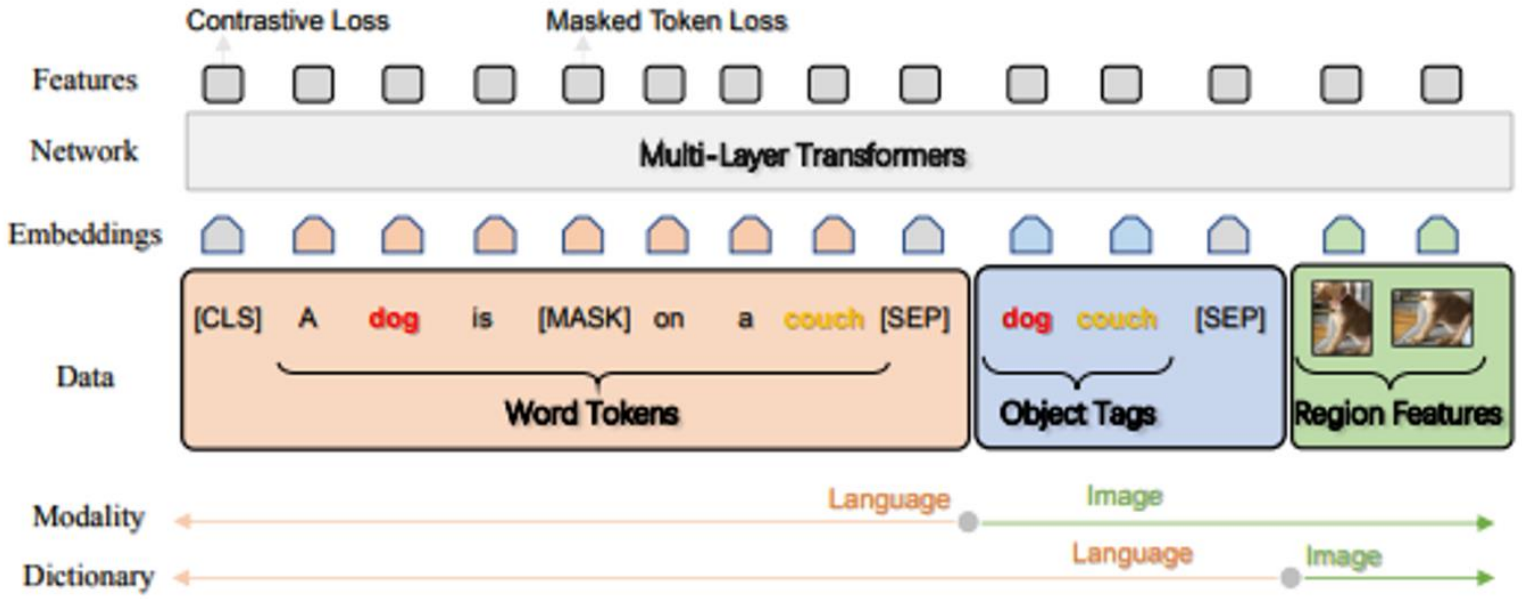

- 650만쌍으로 구성된 대규모 V+L 데이터셋에서 사전훈련
- 7개의 V+L 이해 및 생성 작업에서 fine-tuning
- Anchor point가 VLP를 위한 alignment modeling(정렬 모델링)에서 사용되는 건 처음이라고 함

* Alignment modeling : 두 가지 다른 모달 데이터 관계를 모델링하는 과정(언어-언어, 이미지-텍스트 등)

TRAIN AND TEST

# OSCAR Pre-training

OSCAR : learn representation that capture channel-invariant(or modality-invariant) factors at the semantic level.
-> 의미적인 수준에서 채널의 독립적인 요소를 포착하는 표현을 학습하기 위한 것..
-> 인간은 여러 채널을 통해 세상을 인지한다.. 개별 채널이 불완전하거나 노이지하더라도 여러 채널 간의 공유를 통해 중요한 요소는 여전히 인지 가능하다...




TRAIN AND TEST

# OSCAR Pre-training

## 1) Input

Word-tag-Image 트리플(W,Q,V)로 표현
W : 텍스트의 단어 임베딩 시퀀스
Q : 이미지에서 감지된 객체 태그(텍스트형식)의 단어 임베딩 시퀀스
V: 이미지의 영역 벡터 집합
* Q 도입이 다른 모델과 큰 차이점


Image -> Faster R-CNN -> 객체태그 탐지(Q생성)

↓

Visual semantic(V',Z) ----------→ V
V' : region feature(2048차원) linear projection(W와 같은 차원으로)
Z : region position

TRAIN AND TEST

# OSCAR Pre-training

## 2) Pre-training Objective

OSCAR의 인풋은 두가지 관점으로 볼 수 있음
1) modality view(x) : text와 image 사이의 representation을 구분한다.
2) dictionary view(x′) : 두 개의 다른 semantic space를 구분하는 것

$$\boldsymbol{x} \triangleq [\underbrace{\boldsymbol{w}}_{\text{language}}, \underbrace{\boldsymbol{q}, \boldsymbol{v}}_{\text{image}}] = [\underbrace{\boldsymbol{w}, \boldsymbol{q}}_{\text{language}}, \underbrace{\boldsymbol{v}}_{\text{image}}] \triangleq \boldsymbol{x}'$$

$$\mathcal{L}_{\text{Pre-training}} = \mathcal{L}_{\text{MTL}} + \mathcal{L}_{\text{C}}.$$

TRAIN AND TEST

# OSCAR Pre-training

## 2) Pre-training Objective

1) modality view(x) : text와 image 사이의 representation을 구분한다.

$$h' \triangleq [q, v] \qquad x \triangleq [\underbrace{w}_{\text{language}}, \underbrace{q, v}_{\text{image}}]$$

- q, v 를 이미지 , w를 언어 모달리티로 간주
- q(태그)를 무작위로 샘플링한 다른 태그 시퀀스로 50% 확률로 교체하여 polluted 이미지 표현
- [CLS]의 인코더 출력은 feud vision-language 의 표현으로 (h',w)이고, 마지막에 FC layer을 넣어 이 쌍이 원래 이미지 표현 (y=1)을 포함하는지 또는 오염된 이미지 표현(y=0)을 포함하는지를 예측.

TRAIN AND TEST

# OSCAR Pre-training

## 2) Pre-training Objective

2) dictionary view(x') : 두 개의 다른 semantic space를 구분하는 것

$$h \triangleq [w, q] \qquad [\underbrace{w, q}_{\text{language}}, \underbrace{v}_{\text{image}}] \triangleq x'$$

- q,w를 언어, v를 이미지 모달로 간주
- 사전 훈련을 위해 Masked Token Loss(MTL)사용
- H의 각 입력 토큰을 15%의 확률로 무작위로 가리고, 가려진 토큰 h_i를 [mask] 토큰으로 대체
- 모든 이미지 특징 v를 기반으로 Negative log-likelihood 최소화

$$\mathcal{L}_{\mathrm{MTL}} = -\mathbb{E}_{(\boldsymbol{v},\boldsymbol{h})\sim\mathcal{D}} \log p(h_i|\boldsymbol{h}_{\setminus i}, \boldsymbol{v})$$

TRAIN AND TEST

# OSCAR Pre-training

**3) pre-training Corpus/ Implementation Details**

- 기존 V+L 데이터셋을 기반으로 사전 훈련 말뭉치 구축
  (COCO, Conceptual Captions, SBU captions, filcker, QA)
- 총 410만개의 고유 이미지, 말뭉치는 650만개의 (텍스트-태그-이미지) 트리플

* OSCAR B, OSCAR L -> 각각 BERT base(h=768), BERT large(h=1024)
파라미터는 Bert의 기본 파라미터와, linear projection을 위한 행렬 W

$$\theta = \{\theta_{\text{BERT}}, \mathbf{W}\}$$

*Optimizer Adamw.
Oscar B는 최소 100만 단계까지 학습. lr=5e-5, batch_size=768
OscarL은 최소 90만 단계까지 학습, lr=1e-5, batch_size=512

TRAIN AND TEST

# Experiments

| Task | Image Retrieval | | | Text Retrieval | | | Image Captioning | | | | NoCaps | | VQA | NLVR2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | B@4 | M | C | S | C | S | test-std | test-P |
| SoTA$_S$ | 39.2 | 68.0 | 81.3 | 56.6 | 84.5 | 92.0 | 38.9 | 29.2 | 129.8 | 22.4 | 61.5 | 9.2 | 70.90 | 53.50 |
| SoTA$_B$ | 48.4 | 76.7 | 85.9 | 63.3 | 87.0 | 93.1 | 39.5 | 29.3 | 129.3 | 23.2 | 73.1 | 11.2 | 72.54 | 78.87 |
| SoTA$_L$ | 51.7 | 78.4 | 86.9 | 66.6 | 89.4 | 94.3 | – | – | – | – | – | – | 73.40 | 79.50 |
| OSCAR$_B$ | **54.0** | **80.8** | **88.5** | **70.0** | **91.1** | **95.5** | **40.5** | **29.7** | **137.6** | 22.8 | **78.8** | **11.7** | **73.44** | 78.36 |
| OSCAR$_L$ | **57.5** | **82.8** | **89.8** | **73.5** | **92.2** | **96.0** | **41.7** | **30.6** | **140.0** | **24.5** | **80.9** | **11.3** | **73.82** | **80.37** |
| $\Delta$ | **5.8** ↑ | **4.4** ↑ | **2.9** ↑ | **6.9** ↑ | **2.8** ↑ | **1.7** ↑ | **2.2** ↑ | **1.3** ↑ | **10.7** ↑ | **1.3** ↑ | **7.8** ↑ | **0.5** ↑ | **0.42** ↑ | **0.87** ↑ |

- S는 트랜스포머 based 이전 VLP 모델
- B는 BERT base와 비슷한 크기의 VLP
- L은 BERT large와 비슷한 크기의 VLP

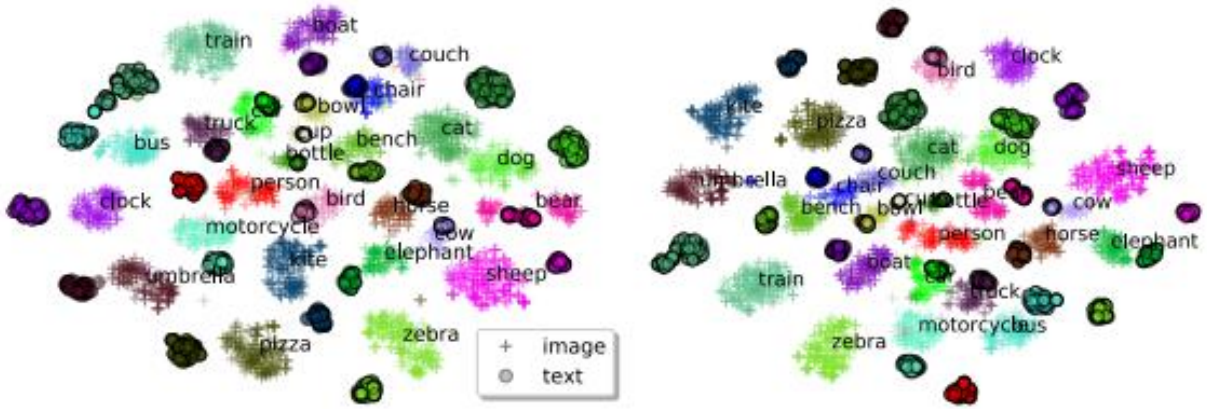
*Image Captioning : 이미지 내용에 대한 텍스트 생성
*NoCaps : Novel object Captioning
*VQA : 이미지 기반으로 자연어 질문에 대답 요구
* NLVR2 : 이미지 쌍과 텍스트를 취하여, 자연어 진술이 이미지 쌍에 대해 참인지 여부를 결정

* B@4 : BLEU-4(4-gram BLEU)
* M : METEOR (비교 일치)
* C : CIDEr(Consensus-based Image Description Evaluation) 단어 중요성으로 유사성 측정
* S : SPICE(Semantic Propositional Image Caption Evaluation) 의미론적 일치 평가

TRAIN AND TEST

# Experiments

## Qualitative Studies




(a) OSCAR (b) Baseline (No tags)

- T-SNE를 이용해 COCO 테스트 세트의 이미지-텍스트 쌍의 학습된 의미적 특성 공간을 2D로 시각화
- 각 이미지 영역과 단어 토큰에 대해 모델을 통과시키고, 마지막 레이어 출력을 사용

1) Oscar을 사용하면 두 모달 간 동일한 객체의 거리가 상당히 줄어듬 ex) 사람, 얼룩말
2) Oscar을 사용하면, 관련 의미의 객체 클래스들이 더 가까워짐.

-> 정렬학습에서 객체 태그의 중요성

TRAIN AND TEST


TRAIN AND TEST