# VLMixer

Unpaired Vision-Language Pre-training via CMC

Mose Park

Department of Statistical Data Science
University of Seoul

Selective. Lab

May 23, 2024

# Index

# Main concepts

Vision-Language
Data Augmentation

Modality Gap

Loss function

# 1

# Problem

# Problem

## Paired

* image

train traveling down a track in front of road

* text

## unpaired

* image

?

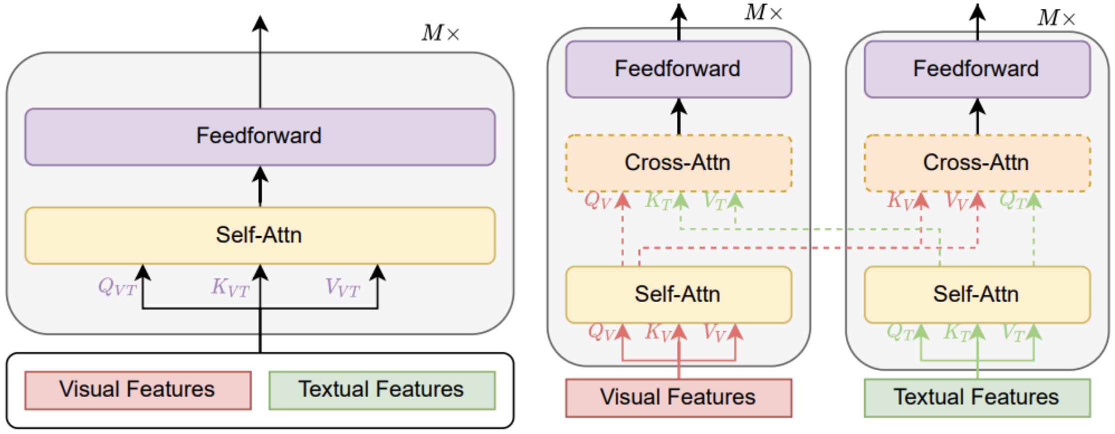
- A train traveling down tracks next to lights.
- A blue train is next to a sidewalk on the rails.
- A passenger train pulls into a train station.

* text

▶ Obtaining paired image-text data is a costly and precise task. → Need data augmentation!

# Paired VLP

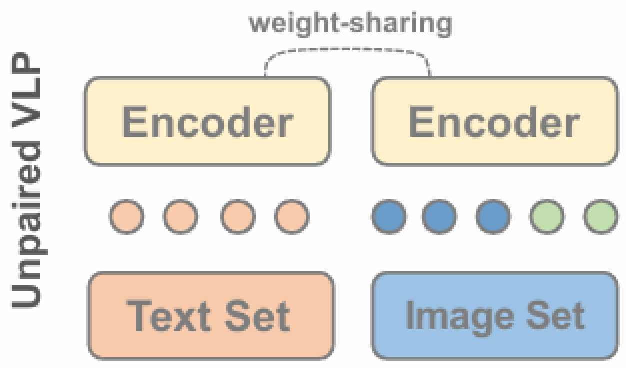
Learning Obj
- single stream
- dual stream




(a) Single-Stream Architecture (b) Dual-Stream Architecture

**Fig. 1** Illustration of two types of model architectures for VLP.

# Unpaired VLP




C. U-VisualBERT

* Same system, Sharing parameter
* Independent dataset

▶ How does this paper address the modality gap?

* VLP: A Survey on Vision-Language Pre-training

# 2

# Approach

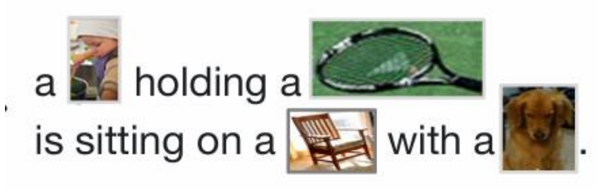
# Approach : Cross-modal CutMix


* Image-only dataset

concept detector

* patch gallery

CutMix




* Cross-modal view **S**

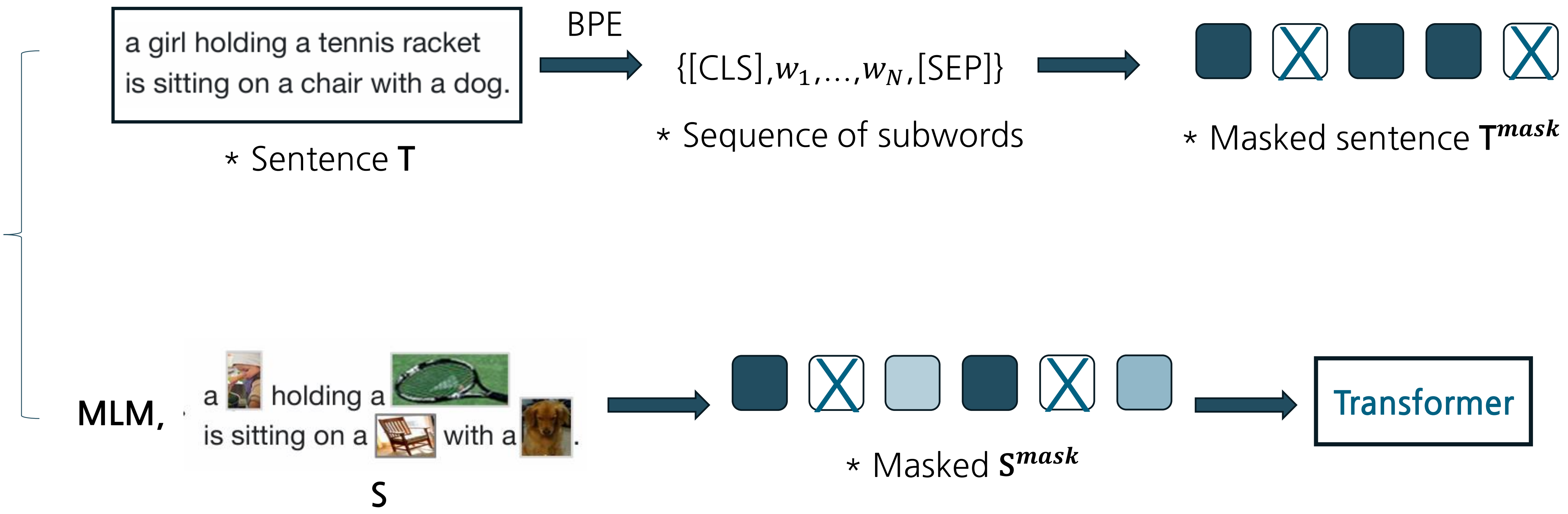
a girl holding a tennis racket
is sitting on a chair with a dog.

* Text dataset

▶ Preserve syntactic and semantic information

▶ Introduce visual tokens as the cross-modality noise

# Approach : VALP




$$\mathcal{L}_{\mathrm{mlm}} = -\mathbb{E}_{\mathbf{T}\sim\mathcal{D}^T} \log(\mathbf{T}|\mathbf{S}^{\mathrm{mask}}) = \mathrm{CE}(\hat{\mathbf{S}}, \mathbf{T})$$

- **Masked language model** exploits multi-modal fusion.
- **Contrastive learning** is to learn cross-modal alignments.

# Contd.

$S_m$ : Positive instance

$\{(T_1, S_1), \cdots, (T_M, S_M)\}$ ⟹ Selecting anchor instance $T_m$

Otherwise : Negative instance

→ Contrastive Learning ! (By Eqn (4) Loss)

$$\mathcal{L}_{\text{cl}} = -\sum_{m=1}^{M} \log \frac{\exp\left(f\left(\mathbf{T}_m^{\text{mask}}, \mathbf{S}_m^{\text{mask}}\right)/\tau\right)}{\sum_{l=1}^{M} \exp\left(f\left(\mathbf{T}_m^{\text{mask}}, \mathbf{S}_l^{\text{mask}}\right)/\tau\right)}$$

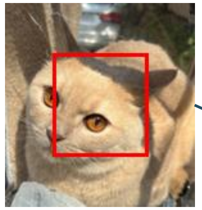
- **Masked language model** exploits multi-modal fusion.
- **Contrastive learning** is to learn cross-modal alignments.

# Approach : TAVP

* Image set $D^I$

* Image $I$

PCD



* region

cat

* tag : Det(**I**)

concat

$$\mathbf{Q} = (\mathbf{I}, Det(\mathbf{I}))$$

* Multi-modal representation

masking tag

$$\mathbf{Q}^{mask} = (\mathbf{I}, Det(\mathbf{I}))$$

* Masked $\mathbf{Q}^{mask}$

Transformer

▶ Effectively combining image and text information using tags from image $I$

# 3

## Details

# Details : Cross-modal CutMix

Faster R-CNN → Detecing important region


* Image-only dataset

concept detector

* patch gallery → $G = \{(w_i^{con}, c_i^{con}, \{w_{ij}^{ctx}, c_{ij}^{ctx}\})\}$

ex) **salient region**

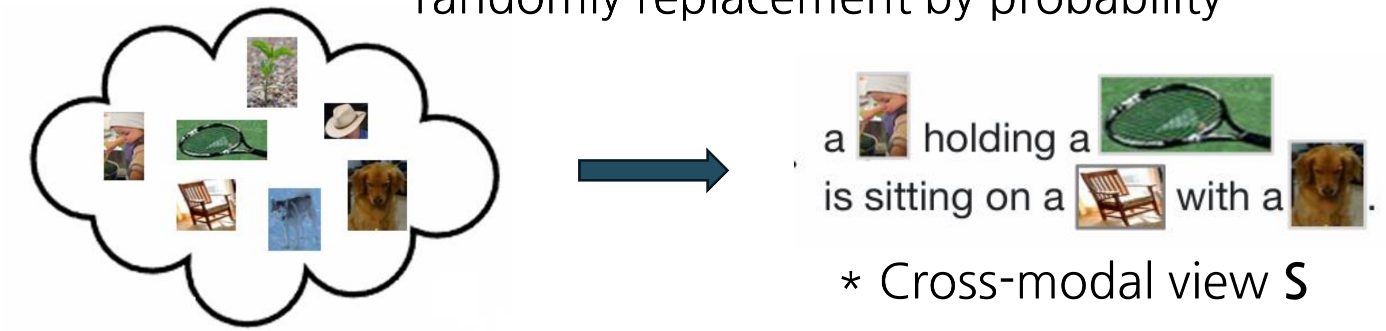
$w_i^{con}$ : dog (concept)

$w_{ij}^{ctx}$ : bench, tree, etc (contextual concept)

▶ Faster R-CNN : Bottom up attention mechanism

▶ w is concept label and c is confidence score.

randomly replacement by probability




* Cross-modal view **S**

* patch gallery

Control the importance of contextual concepts

$$p_i = \begin{cases} c_i^{\text{con}} + \frac{r_{\text{ctx}}}{|G_i|} \sum_{w_{ij}^{\text{ctx}} \in G_i} c_{ij}^{\text{ctx}}, & \text{if } w_i^{\text{con}} = w_n \\ 0, & \text{otherwise} \end{cases}$$

The # of set of contextual concepts related to the word

▶ The text is replaced with patches considering the contextual situation.

# Details : VALP - Byte pair encoding

a girl holding a tennis racket is sitting on a chair with a dog.

⋆ Sentence **T**

$\xrightarrow{\text{BPE}}$ $\{[\text{CLS}], w_1, \ldots, w_N, [\text{SEP}]\}$

⋆ Sequence of subwords

⋆ Masked sentence $\mathbf{T}^{mask}$

ex)

$$\{[CLS], lower, high, [SEP]\} \longrightarrow \{[CLS], low, er, hi, gh, [SEP]\}$$

▶ BPE : word to subword (original → small)

▶ I think that since patches are smaller than original image, also words will be sliced.

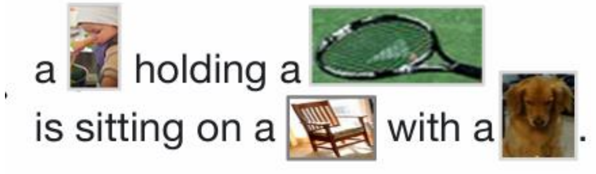
# Details : VALP - Constrative Learning

▶ $$\mathcal{L}_{\mathrm{cl}} = -\sum_{m=1}^{M} \log \frac{\exp\left(f\left(\mathbf{T}_m^{\mathrm{mask}}, \mathbf{S}_m^{\mathrm{mask}}\right)/\tau\right)}{\sum_{l=1}^{M} \exp\left(f\left(\mathbf{T}_m^{\mathrm{mask}}, \mathbf{S}_l^{\mathrm{mask}}\right)/\tau\right)}$$

$f(\ )$

Toy example)

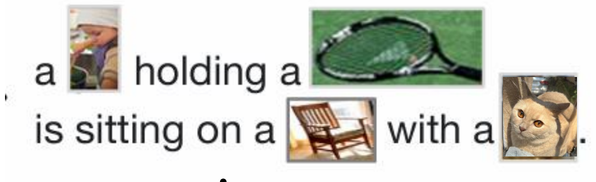
distance ↓↓↓




a girl holding tennis racket
is sitting on a chair with a dog.

*** anchor instance**

**+ instance**




a man holding tennis racket
is sitting on a chair with a cat.

**- instance**

distance ↑↑↑

▶ I show that not masked example for explanation.

# 4

# Experiment

# Datasets

| Dataset | Images | Texts | Text Domain |
|---|---|---|---|
| COCO (train) | 112K | 560K | Image Caption |
| Conceptual Captions (train) | 3M | 3M | Image Caption |
| SBU Caption (all) | 840K | 840K | Image Caption |
| Flickr30k (train) | 29K | 145K | Image Caption |
| VQA (train) | 83K | 445K | Question |
| GQA (train) | 79K | 1.0M | Question |
| VG-QA (train) | 87K | 931K | Question |
| MSVD (train) | - | 48K | Video Caption |
| MSRVTT (train) | - | 130K | Video Caption |
| VATEX (train) | - | 260K | Video Caption |
| ActivityNet Captions (train) | - | 36K | Video Caption |
| Shutterstock (all) | - | 1M | Caption |
| BookCorpus | - | 14M | General Text |
| OpenImages (od train) | 1.67M | - | - |

- Image and text datasets

# Setting

❖ Backbone Model

- Base Transformer
  - 12 layers of transformer blocks
  - Hidden size :768

❖ Position Embedding

- Language/Tag tokens
  - Learning position embedding
- Patch/Image Tokens
  - Linear projection of spatials positions

❖ Restriction of token length

- TAVP : 100
- VALP : 80

❖ Fine tuning

- VLMixer : BERT - Base
- image and text 300k pre-trained
- lr : 5e-5, mini batch size : 1024
- optim : adam

❖ ETC.

- patch detector : Resnet-152 C4
- 15-shot CMC
- $r_{cmc}$ : 0.5
- $r_{ctx}$ : 0.5
- CMCL : temperature ratio : 0.1

| Method | Pre-training Data | | VQA | NLVR$^2$ | | Text Retrieval | | | Image Retrieval | | | GQA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Image | Text | Test-Dev | Dev | Test | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | Test-Dev |
| **Paired VLP** | | | | | | | | | | | | |
| UnicoderVL$_{\mathrm{base}}$ (Li et al., 2019a) | | | - | - | - | 62.3 | 87.1 | 92.8 | 46.7 | 76.0 | 85.3 | - |
| UNITER$_{\mathrm{base}}$ (Chen et al., 2019) | | | 72.27 | 77.14 | 77.87 | 63.3 | 87.0 | 93.1 | 48.4 | 76.7 | 85.9 | - |
| OSCAR$_{\mathrm{base}}$ (Li et al., 2020b) | | | 73.16 | 78.07 | 78.36 | 70.0 | 91.1 | 95.5 | 54.0 | 80.8 | 88.5 | 61.58 |
| VILT$_{\mathrm{base}}$ (Kim et al., 2021) | | | 71.26 | 75.70 | 76.13 | 61.5 | 86.3 | 92.7 | 42.7 | 72.9 | 83.1 | - |
| VinVL$_{\mathrm{base}}$ (Zhang et al., 2021) | | | 75.95 | 82.05 | 83.08 | 74.6 | 92.6 | 96.3 | 58.1 | 83.2 | 90.1 | 65.05 |
| ALBEF (Li et al., 2021a) | | | 75.84 | 82.55 | 83.14 | 77.6 | 94.3 | 97.2 | 60.7 | 84.3 | 90.5 | - |
| **Unpaired VLP** | | | | | | | | | | | | |
| BERT$_{\mathrm{base}}$ (Devlin et al., 2019) | None | None | 64.85 | 51.30 | 51.34 | 57.44 | 84.00 | 91.58 | 44.03 | 74.12 | 84.06 | 50.20 |
| VinVL$_{\mathrm{unpaired}}$ (Zhang et al., 2021) | COCO | COCO | 71.78 | 71.14 | 72.01 | 61.92 | 86.90 | 93.08 | 46.90 | 76.18 | 85.53 | 62.24 |
| U-VisualBERT (Li et al., 2021b)* | COCO | COCO | 72.41 | - | - | - | - | - | - | - | - | - |
| VLMixer | COCO | COCO | **72.60** | **72.71** | **73.08** | **62.69** | **87.35** | **93.64** | **47.95** | **77.06** | **86.22** | **63.13** |
| U-VisualBERT (Li et al., 2021b) | CC3M | CC3M+BC | 70.74 | 71.74 | 71.02 | - | - | - | - | - | - | - |
| VinVL$_{\mathrm{unpaired}}$ (Zhang et al., 2021) | CC3M | CC3M | 72.20 | 68.96 | 68.94 | 62.08 | 86.04 | 93.00 | 47.29 | 76.15 | 85.53 | 63.12 |
| VLMixer | CC3M | CC3M | 72.66 | 74.31 | 73.86 | 62.20 | 86.32 | 92.80 | 47.44 | 76.22 | 85.41 | 62.65 |
| VLMixer | Full | Full | **72.89** | **76.61** | **77.01** | **64.76** | **88.56** | **94.22** | **50.06** | **78.36** | **86.91** | **63.25** |