MASS: Overcoming Language Bias in Image-Text Matching

Jiwan Chung[1], **Seungwon Lim**[1], **Sangkyu Lee**[1], **Youngjae Yu**[1]

[1]Yonsei University
50 Yonsei-ro, Seodaemun-gu
Seoul, South Korea
jiwan.chung.research@gmail.com

- Problem / objective
  - Language bias in image-text matching
    - Models predominantly rely on language priors and neglect to adequately consider the visual content
- Contribution / Key idea
  - Multimodal ASsociation Score (**MASS**)
    - A framework that reduces the reliance on language priors for better visual accuracy in image-text matching problems
    - Inference-time framework (i.e., training-free)

- **Motivation**

  - ❏ VLM (e.g., CLIP): Image-text matching 잘해, but, 정확한 language modeling은 못해
    i) Linguistic compositionality
    ii) Linguistic constructs (e.g., existence, quantity)
  - ❏ Better understanding of linguistic structures by objective change
    - Contrastive objective in CLIP → Log-likelihood induced from the autoregressive image captioning objective
  - ❏ Caveat in directly using the image captioning models to assess image-text similarity
    - Language bias: the propensity of VLMs to rely heavily on language priors in the training data, instead of properly conditioning their output on the given images
  - ❏ Thus, we propose Multimodal ASsociation Score (MASS) as an inference-time framework designed to reduce language bias in image-text matching

Figure 1: Captions retrieved with each method given the image, where only MASS succeeds in ruling out the failure modes. Models trained Image-Text Contrastive (ITC) objectives such as CLIP (Radford et al. 2021) fail to model linguistic structure. Token Likelihood (TL) of image captioning models including OFA (Wang et al. 2022b) shows over-reliance on its language prior. Our MASS amends the language bias of image captioning models for accurate image-text matching capability.

- **Image-Text Similarity**

1. Divide the similarity functions by the granularity of the training objectives
2. **Sequence-level Similarity:** Image-Text Contrastive Learning (ITC), Image-Text Matching (ITM)
   a. ITC
      i. CLIP (vision encoder, text encoder)
      ii. Similarity score:

$$\mathcal{S}_{\text{ITC}}(\mathbf{c}, \mathbf{x}) := \frac{f_{\bar{\phi}}(\mathbf{c})^T \cdot g_{\bar{\psi}}(\mathbf{x})}{||f_{\bar{\phi}}(\mathbf{c})|| \cdot ||g_{\bar{\psi}}(\mathbf{x})||} \quad (1)$$

   b. ITM
      i. ALIGN (multi-modal encoder), linear classifier
      ii. Similarity score:

$$\mathcal{S}_{\text{ITM}}(\mathbf{c}, \mathbf{x}) = \frac{\exp(h_{\bar{\omega}}(f_{\bar{\phi},\bar{\psi}}(\mathbf{c}, \mathbf{x})))}{1 + \exp(h_{\bar{\omega}}(f_{\bar{\phi},\bar{\psi}}(\mathbf{c}, \mathbf{x})))} \quad (2)$$

3. **Token-level Similarity:** Token-Level Supservision (TL)
   a. TL
      i. Image captioning model
      ii. Token-level similarity score: Token-level likelihood of an image captioning model $p_{\bar{\theta}}(x_t|x_{<t}, \mathbf{c})$
      iii. Sequence-level similarity score: Mean of all text token scores in the sequence

$$\mathcal{S}_{\text{TL}}(\mathbf{c}, \mathbf{x}) := \frac{1}{l} \sum_{t<l} p_{\bar{\theta}}(x_t|x_{<t}, \mathbf{c}) \quad (3)$$

- **Motivation**

  ❏ Problem of directly using likelihood from autoregressive VLMs as an image-text similarity function
  - Not a pure image-text similarity
  - ***Linguistically plausible captions*** are assigned a high likelihood.

  ❏ Solution: ***Pointwise mutual information (PMI)***
  - Utilized in NLP tasks but not in VLM
  - So, we propose ***PMI*** as an effective method for reducing language bias in image-text matching

$$PMI(\mathbf{x};\mathbf{c}) := \log \frac{p(\mathbf{x},\mathbf{c})}{p(\mathbf{x})p(\mathbf{c})} = \log \frac{p(\mathbf{x}|\mathbf{c})}{p(\mathbf{x})} \qquad (4)$$

- **Multimodal ASsociation Score (MASS)**

- Average of each token's pointwise mutual information over the total length l.

$$\log p_{\bar{\theta}}(\mathbf{x}|\mathbf{c}) = \underbrace{\log p_{\bar{\theta}}(\mathbf{x})}_{\text{linguistic}} + \underbrace{\log \frac{p_{\bar{\theta}}(\mathbf{c}|\mathbf{x})}{p_{\bar{\theta}}(\mathbf{c})}}_{\text{association}} \quad (5)$$

Linguistic probability of text **x**

$$\mathcal{S}_{\text{MASS}}(\mathbf{c}, \mathbf{x}) := \frac{1}{l} \sum_{t<l} \log \frac{p_{\bar{\theta}}(x_t|x_{<t}, \mathbf{c})}{p_{\bar{\theta}}(x_t|x_{<t})} \quad (6)$$

- Token-level similarity:
각 토큰마다 token-level log-likelihood에서 token-level linguistic log-likelihood를 뺌

$$\log p_{\bar{\theta}}(x_t|x_{<t}) = \int_{\mathcal{C}} \log p_{\bar{\theta}}(x_t|x_{<t}, \mathbf{c}) d\mathbf{c} \quad (7)$$

$$\approx \frac{1}{N} \sum_{i=1} \log p_{\bar{\theta}}(x_t|x_{<t}, \tilde{\mathbf{c}}_i) \quad (8)$$

- Autoregressive VLM의 이미지 인풋 어떡할건지
i) 식7,8: Monte Carlo approximation -> computation issue로 별로
ii) 식9: 인풋으로 null image 사용
-> 식9가 최종.

$$\mathcal{S}_{\text{MASS}}(\mathbf{c}, \mathbf{x}) \approx \frac{1}{l} \sum_{t<l} \log \frac{p_{\bar{\theta}}(x_t|x_{<t}, \mathbf{c})}{p_{\bar{\theta}}(x_t|x_{<t}, \mathbf{c}_\emptyset)} \quad (9)$$