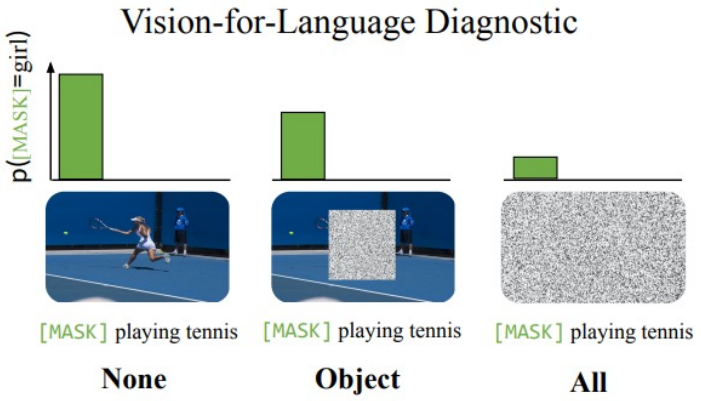
Vision-*and*-Language or Vision-*for*-Language? On Cross-Modal Influence in Multimodal Transformers

Stella Frank[*,ð] **Emanuele Bugliarello**[*,ɛ] **Desmond Elliott**[ɛ]
[ð]University of Trento [ɛ]University of Copenhagen
stella.frank@unitn.it {emanuele, de}@di.ku.dk

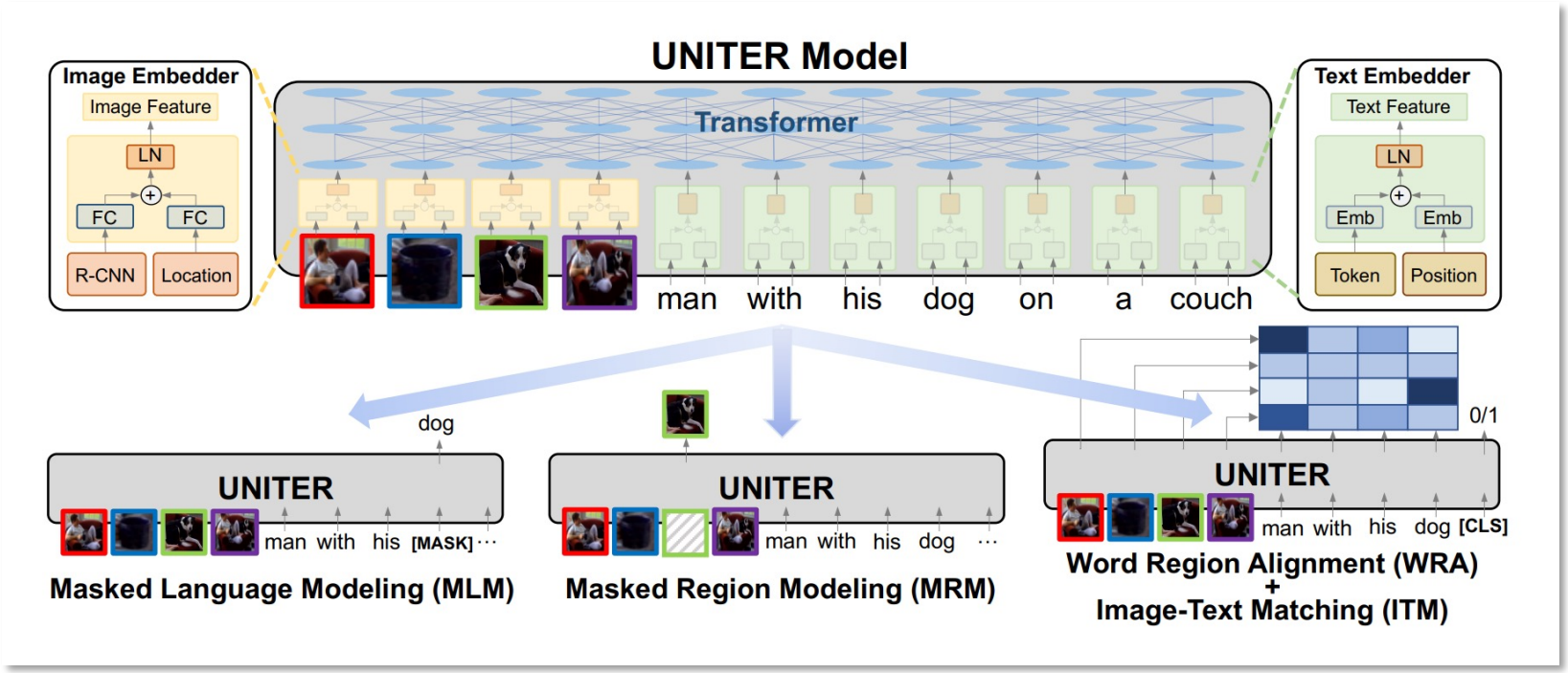
Abstract

Pretrained vision-and-language BERTs aim to learn representations that combine information from both modalities. We propose a diagnostic method based on *cross-modal input ablation* to assess the extent to which these models actually integrate cross-modal information. This method involves ablating inputs from one modality, either entirely or selectively based on cross-modal grounding align-



Vision-*and*-Language or Vision-*for*-Language?
On Cross-Modal Influence in Multimodal Transformers

- Recent surge of multi-modal models
- Vision and Language variants leverage transformers, BERT-like

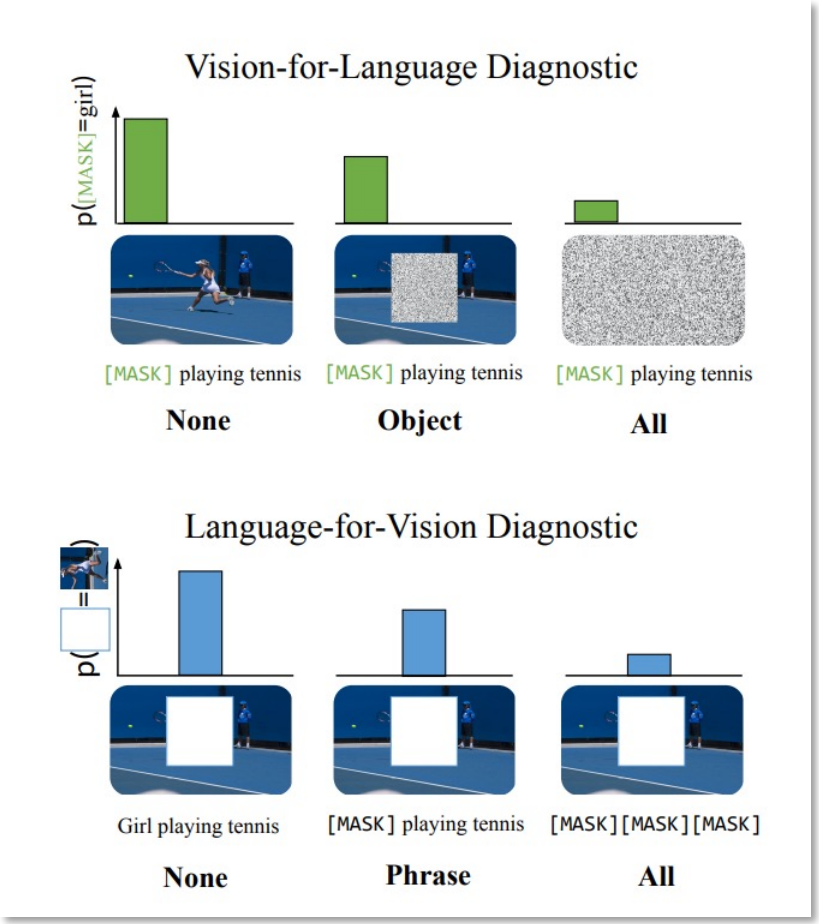


UNITER: Learning UNiversal Image-TExt Representations, 2019, Chen et al.

Vision-*and*-Language or Vision-*for*-Language?
On Cross-Modal Influence in Multimodal Transformers

- How do these models *combine information from both modalities*?
- Is it Text prevalent? Or Vision?

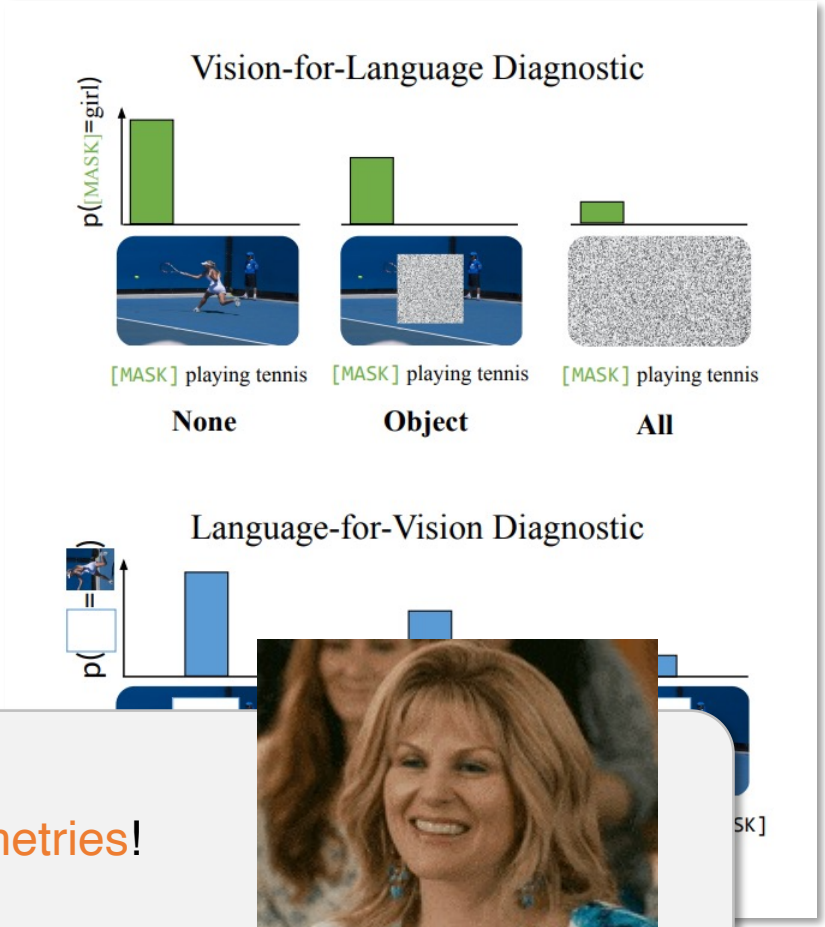
- The authors proposed a cross-modal input ablation diagnostic method
- …or
- *How good are models are predicting text, if vision is ablated*?
- *How good are models are predicting vision, if text is ablated*?



Vision-*and*-Language or Vision-*for*-Language?
On Cross-Modal Influence in Multimodal Transformers

- How do these models *combine information from both modalities*?
- Is it Text prevalent? Or Vision?

- The authors proposed a cross-modal input ablation diagnostic method
- …or
- *How good are models are predicting text, if vision is ablated*?
- *How good are models are predicting vision, if text is ablated*?

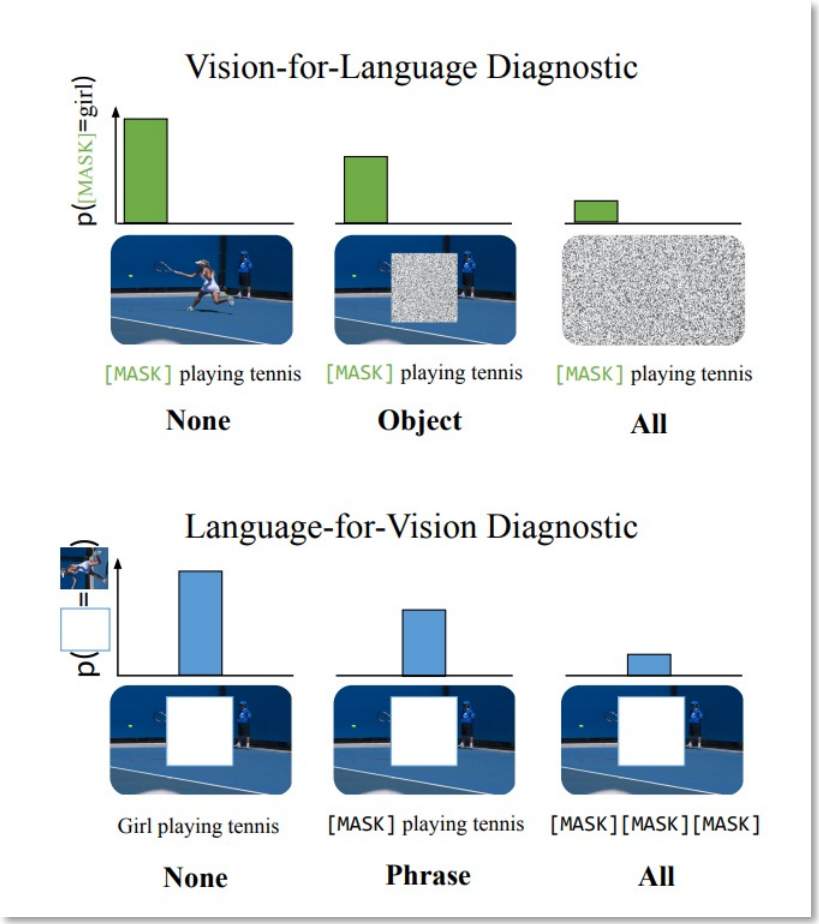


There are asymmetries!

Vision-*and*-Language or Vision-*for*-Language?
On Cross-Modal Influence in Multimodal Transformers

Cross-modal input ablation diagnostic method

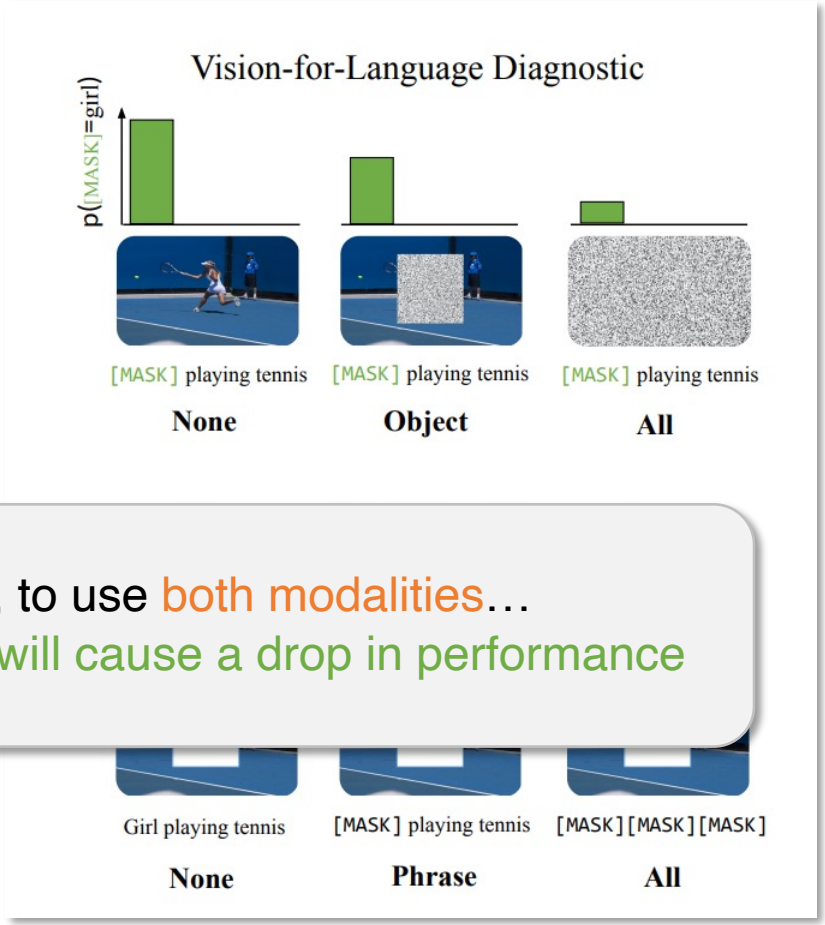
- Straightforward to perform
- Previous work focused on interpreting activations or attention – *difficult*



Vision-*and*-Language or Vision-*for*-Language?
On Cross-Modal Influence in Multimodal Transformers

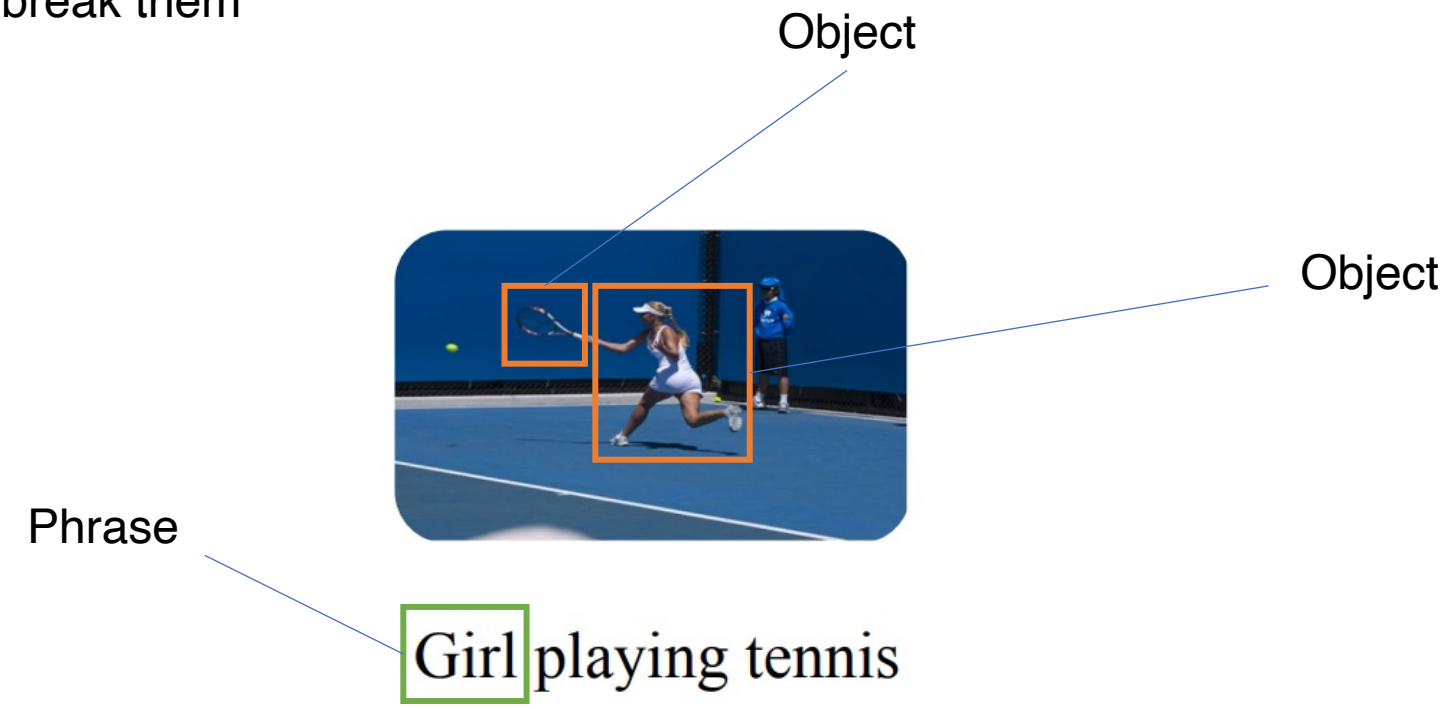
Cross-modal input ablation diagnostic method

- Straightforward to perform
- Previous work focused on interpreting activations or attention – *difficult*



If the model learns, during training, to use both modalities…
… ablating one of two, during evaluation, will cause a drop in performance

- What should we ablate?
- We suppose that models learn alignments between visual concepts and phrases
- Let's break them



Vision-*and*-Language or Vision-*for*-Language?
On Cross-Modal Influence in Multimodal Transformers

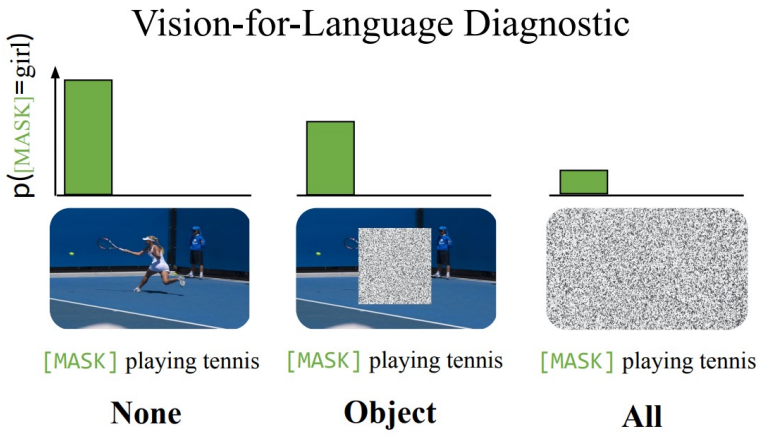
- Vision-for-Language Diagnostic
- Ablate: None, Object, All

- Language task:
- Masked Language Modelling

Masked language modelling The MLM task is to predict the identity of a set of masked tokens $\mathbf{w_m}$, given unmasked tokens $\mathbf{w}_{\backslash \mathbf{m}}$ and visual context $\mathbf{v}$:

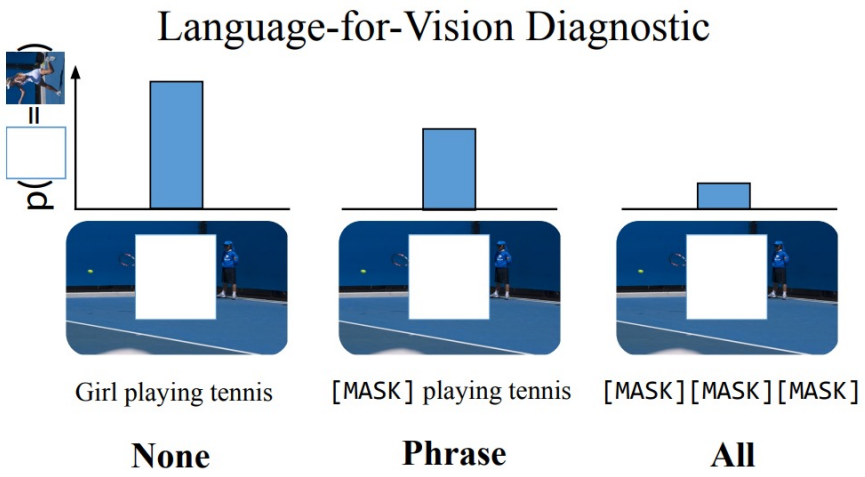
$$\text{MLM}(\mathbf{m}, \mathbf{w}, \mathbf{v}; \theta) = -\sum_{i \in \mathbf{m}} \log_2 \mathbb{P}_\theta(\mathbf{w}_i | \mathbf{w}_{\backslash \mathbf{m}}, \mathbf{v}), \quad (1)$$

where θ denotes a model's parameters.



Vision-*and*-Language or Vision-*for*-Language? On Cross-Modal Influence in Multimodal Transformers

- Language-for-Vision Diagnostic
- Ablate: None, Phrase, All



- Vision task:
- Masked Region Classification

Masked region classification The MRC task is to predict the object class of a masked visual region $\mathbf{v}_i$ given unmasked visual context $\mathbf{v}_{\setminus \mathbf{m}}$ and tokens $\mathbf{w}$. The MRC-KL variant (Li et al., 2019) measures the KL-divergence of the predicted distribution rather than the cross-entropy against a single object class. For each masked region $\mathbf{v}_i$ linked to a phrase, MRC-KL is computed as follows:

$$\text{MRC-KL}(\mathbf{w}, \mathbf{v}_i; \theta) = \text{KL}(\mathbb{P}_g(\mathbf{v}_i) || \mathbb{P}_\theta(\mathbf{v}_i | \mathbf{w}, \mathbf{v}_{\setminus \mathbf{m}})), \quad (2)$$

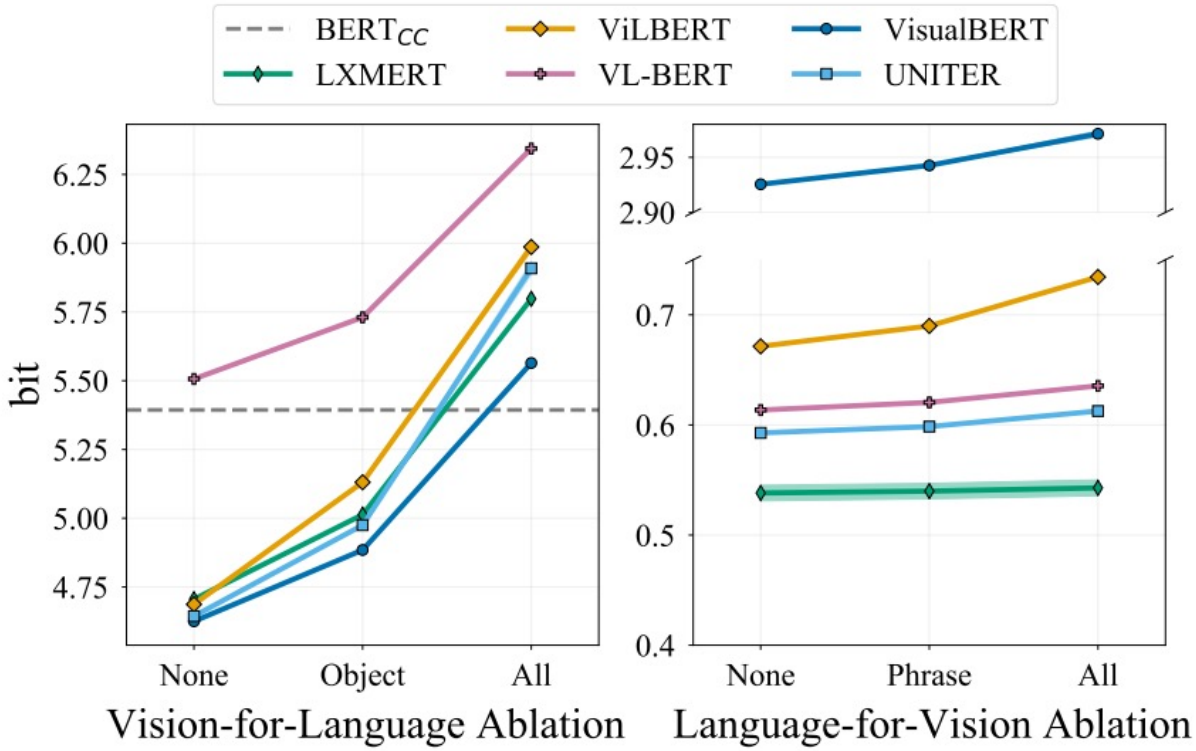
where $\mathbb{P}_g$ is the target object distribution and $\mathbb{P}_\theta$ is the distribution predicted by the model. During

Vision-*and*-Language or Vision-*for*-Language? On Cross-Modal Influence in Multimodal Transformers

- Experimental setting
- Models:
  - LXMERT (Tan and Bansal, 2019), ViL-BERT (Lu et al., 2019) (dual-stream); VLBERT (Su et al., 2020), VisualBERT (Liet al., 2019) and UNITER (Chen et al., 2020) (single-stream)
- Pretraining on Conceptual Captions
  - Objectives: MLM, MRC-KL, Image-Text alignment
- "Silver" labels for MRC provided by Faster R-CNN
  - ~ 1650 classes
- Region overlap: Intersection over Union, Intersection over target

$$\mathrm{IoU}(\mathbf{b}_i, \mathbf{b}_j) = \frac{|\mathbf{b}_i \cap \mathbf{b}_j|}{|\mathbf{b}_i \cup \mathbf{b}_j|}. \qquad \mathrm{IoT}(\mathbf{b}_i, \mathbf{b}_j) = \frac{|\mathbf{b}_i \cap \mathbf{b}_j|}{|\mathbf{b}_i|}.$$

Vision-*and*-Language or Vision-*for*-Language?
On Cross-Modal Influence in Multimodal Transformers

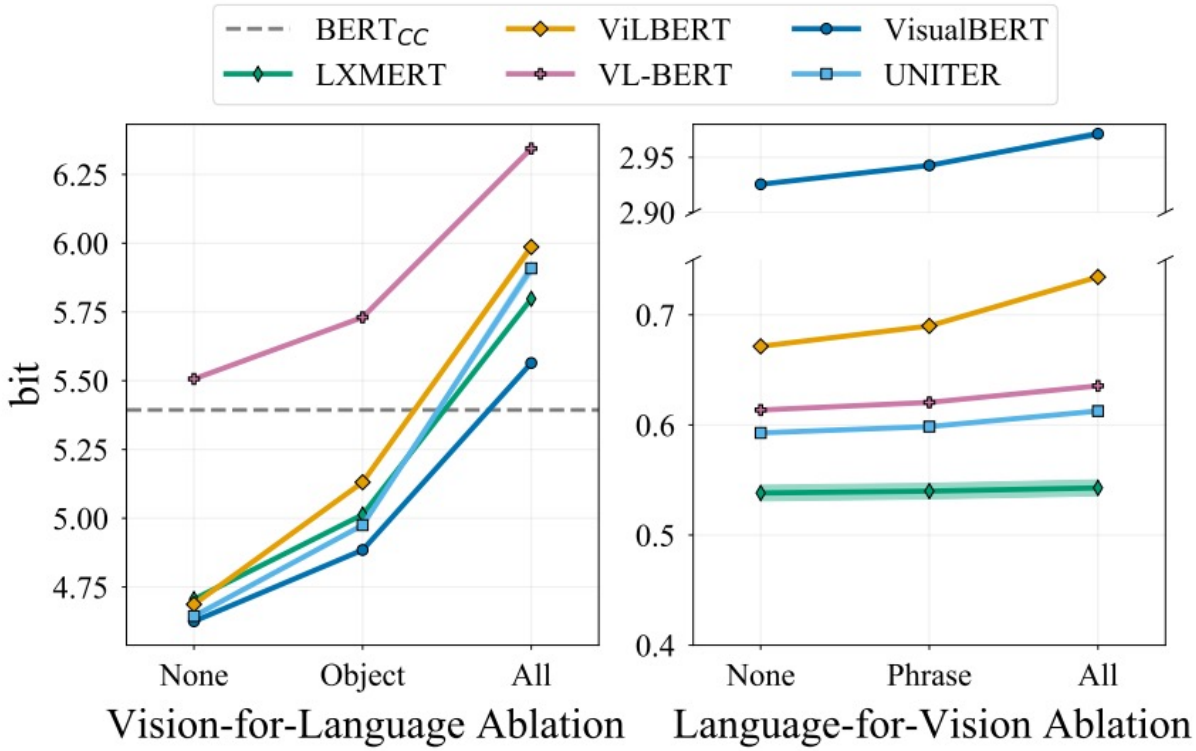


- As expected, ablation removes useful information

Vision-for-Language (good)

The drop removing *All* is much more significant than removing *Object*

Vision-*and*-Language or Vision-*for*-Language? On Cross-Modal Influence in Multimodal Transformers



- As expected, ablation removes useful information

Language-for-Vision (bad)

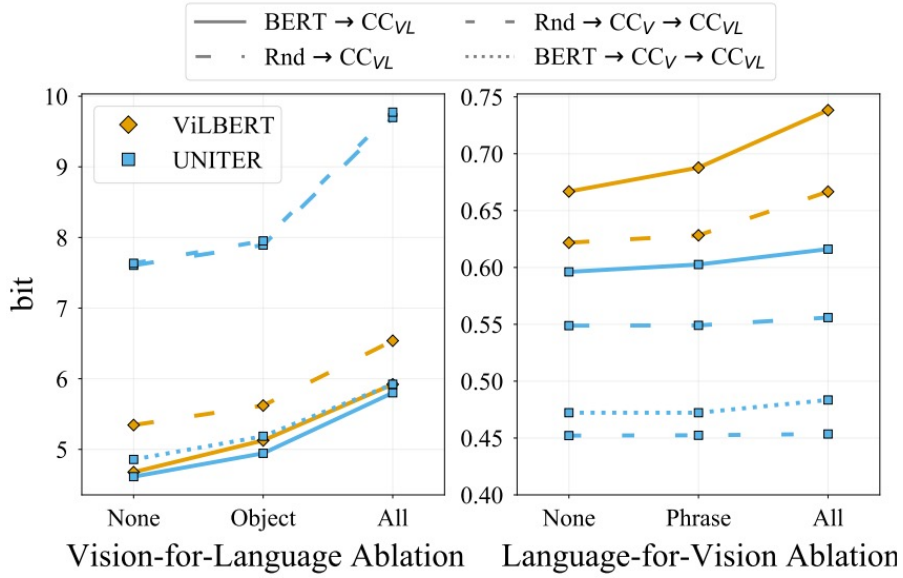
Small 0.5%-3% drop: language for vision is much less used

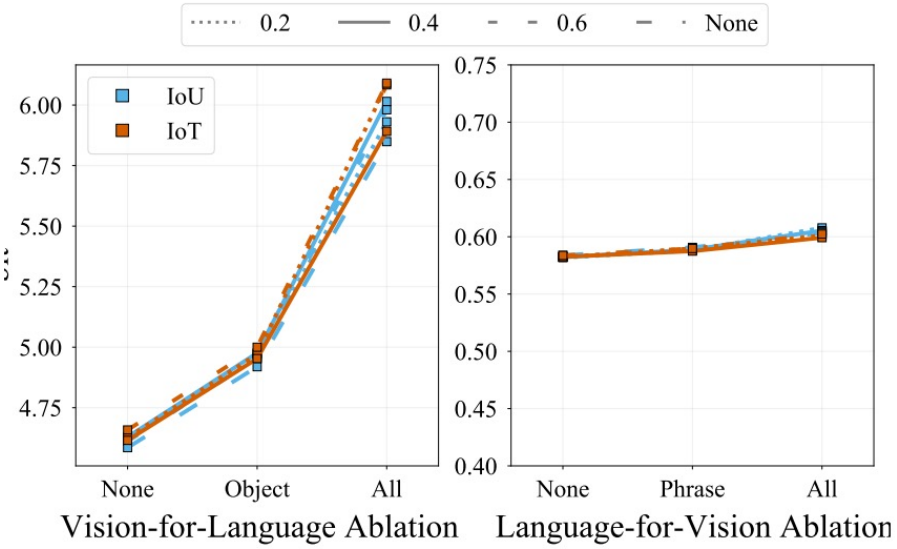
Vision-*and*-Language or Vision-*for*-Language?
On Cross-Modal Influence in Multimodal Transformers

Language-for-Vision (bad)

Small 0.5%-3% drop: language for vision is much less used

- Why is that?
- Different initializations or masking strategies do not impact





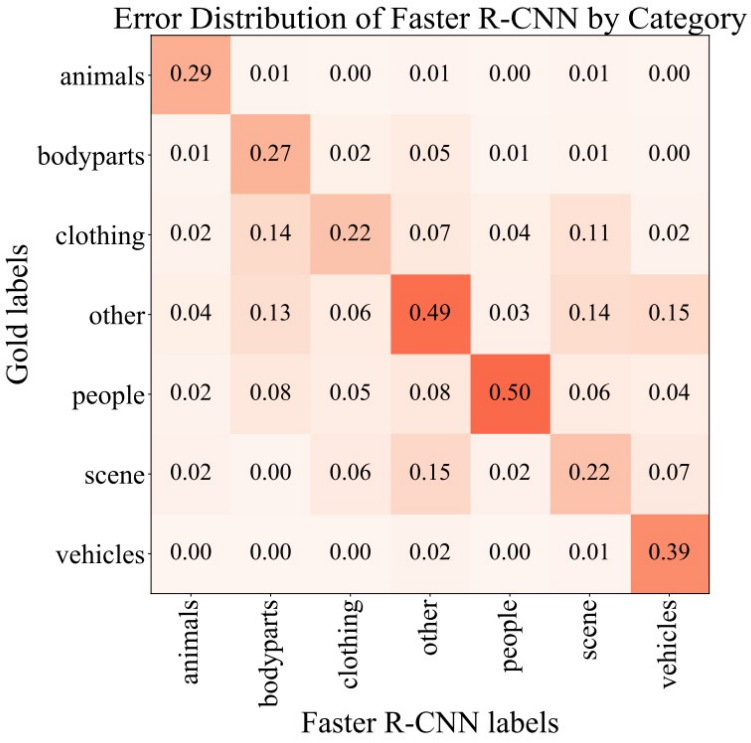
Vision-*and*-Language or Vision-*for*-Language?
On Cross-Modal Influence in Multimodal Transformers

Language-for-Vision (bad)

Small 0.5%-3% drop: language for vision is much less used

- But "silver" labels are noisy
- LabelMatch: pick from Flickr30k Entities records where phrases match with Faster R-CNN categories
- Faster R-CNN is bad at labeling objects

The silver distributions are noisy On the LabelMatch subset, the highest-probability class from Faster R-CNN agrees with the gold label only on 38% of examples. The gold label is in the top-3 55% of the time, in top-5 64%, and in top-10 75% of the time. Figure 6 shows the distribution of errors, grouped into the higher-level categories defined in the Flickr30k Entities dataset. Faster



The authors' take

- Asymmetry in pretrained vision and language models
  - Prediction of Text relies on Vision activations, but not vice versa
- "Silver" labels are unreliable
  - They might introduce more noise than expected if pretraining language-for-vision
- The asymmetry does not affect performance on downstream tasks (sequence classification, visual question answering, etc.) (Bugliarello et al., 2021)
- In the future, we need increased work on language-for-vision tasks (e.g., text-modulated object detection)

Thanks!