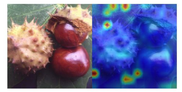
TAB: Transformer Attention Bottlenecks enable User Intervention and Debugging in Vision-Language Models

Pooyan Rahmanzadehgervi, Hung H. Nguyen, Rosanne Liu, Long Mai, Anh Totti Nguyen

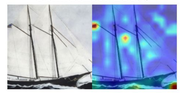
https://aub.ie/tab

Motivation

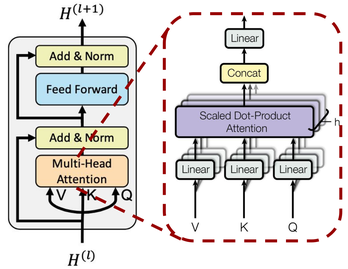
- There is no reliable method for accurately identifying what exact input patches ViTs are attending to.
- There is no causal relation between attention maps and the predictions.
- ViTs often require extra registers for a human-desired explanation.
- There are many heads and layers in ViTs. Thus, it is not trivial to accurately assign credit to each and to perform intervention for prediction verification.



Classification: Chestnut
Confidence: 100%



Classification: Schooner
Confidence: 97%



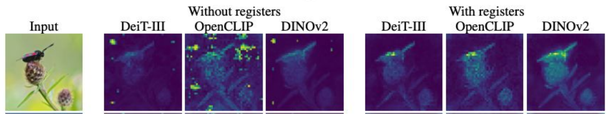
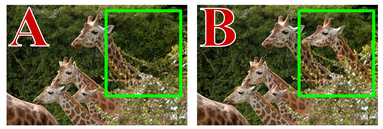


Image Difference Captioning

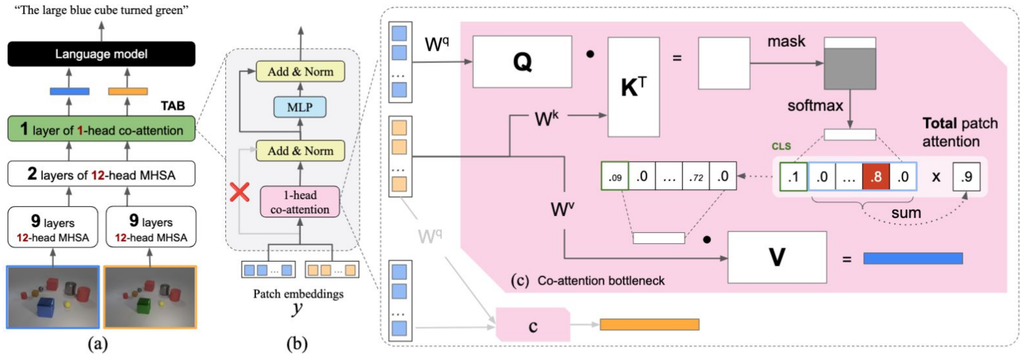
- IDC is a core task behind many real-world applications, e.g., remote sensing, camera surveillance, medical imaging, and urban planning.
- Detecting differences is also an important aspect of many self-supervised methods, e.g., SimCLR.
- Setup:
  - Inputs: (Image **A**, Image **B**)
  - **Output:** A natural language description of potential object-level changes



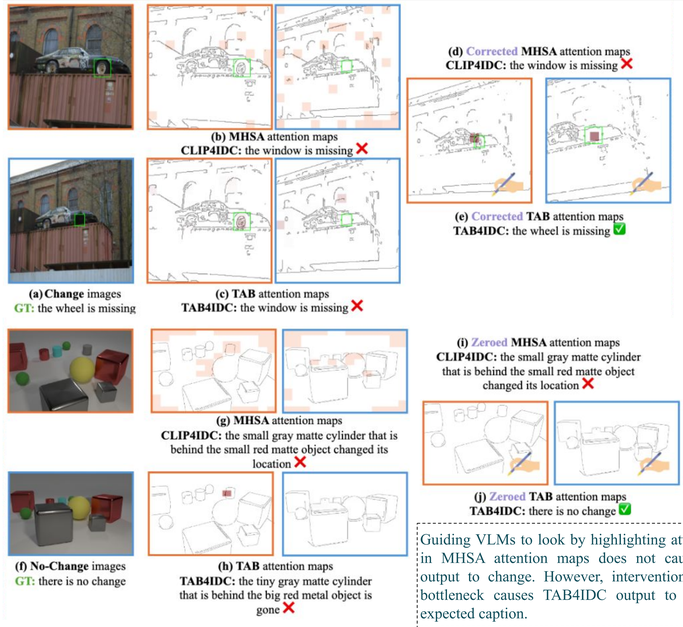
Output: The giraffe has been newly appeared

TAB4IDC

- We apply the following **architectural changes** to the MHSA and the VLM:
  - Replacing cross-attention by **co-attention**.
  - Reducing the **number of heads** from 12 to 1.
  - Removing the initial **skip connection**.
  - Adding a **dynamic attention gating** mechanism that zeroes out the [CLS] attention when the total attention over image patches is zero.
- We train the VLM in **2 stages:**
  1. **Adaptation:** adapting the visual and textual representations via contrastive retrieval loss.
  2. **Captioning:** we supervise the VLM to generate a caption given two images using a Cross Entropy loss and an attention supervision loss.



Results



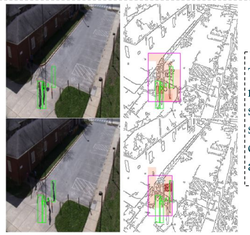
Guiding VLMs to look by highlighting attention patches in MHSA attention maps does not cause CLIP4IDC output to change. However, intervention on the TAB bottleneck causes TAB4IDC output to change to an expected caption.

Method	ViT	Attn. Sup.	B-4	M	R	C	BERTScore
CLIP4IDC [27]	B/32	✗	82.8	56.0	93.6	296.5	92.4
CLIP4IDC [27]	B/16	✗	87.5	60.3	95.6	328.7	95.1
TAB4IDC	B/32	✓	91.4	62.1	94.7	302.0	93.8 (+1.4)
TAB4IDC	B/16	✓	**92.4**	**64.5**	**96.9**	**335.6**	**96.6** (+1.5)
TAB4IDC	B/16	✗	91.9	63.8	96.6	324.4	96.2 (+1.1)

Our proposed TAB4IDC outperforms CLIP4IDC on captioning changes across natural images.

Method	Train	Thresh.	Change		No-change		Mean	
CYWS [61]		✗	100.0	99.91	0.0	0.0	50.0	49.95
CYWS [61]		✓	99.92	81.73	100.0	99.72	**99.96**	89.76
CLIP4IDC [27]		✓	74.9	24.55	12.6	83.74	43.75	54.14
TAB		✓	75.9	98.40	100.0	99.98	87.95 (+44.2)	**99.19**

TAB substantially outperforms MHSA layers in zero-shot chang localization under PG'.



The attention maps in TAB can localize multiple changes in real-world surveillance datasets. Yet, the values spread over the patches, including the changed objects, which leads to lower attention values.

Limitations

- TAB requires the attention supervision loss to achieve its superior localization performance on change pairs.
- The attention map in TAB requires more nonzero values over image patches for multi-change images, which makes the softmax function spread over more patches.
- For the smaller changes, TAB needs smaller patches for a more accurate localization and better captioning performance.

Conclusion

- TAB is a self-explainable and editable bottleneck layer with a 1-head attention for IDC.
- TAB enables an interactive interface allows users to intervene in decision-making, by which one can correct and audit VLMs' decisions.
- Users can use TAB to evaluate how the attended patches in an attention map are important to VLM predictions.