It is not enough adding features unchecked, trusting they are used by the Transformer

Probing the Role of Positional Information in Vision-Language Models

Philipp J. Rösch *philipp.roesch@unibw.de*
Jindřich Libovický *libovicky@ufal.mff.cuni.cz*

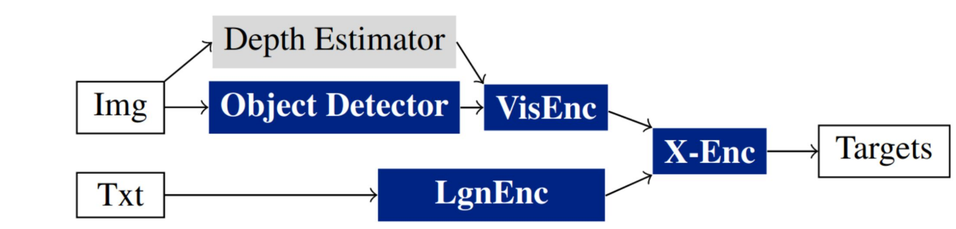
der Bundeswehr Universität München

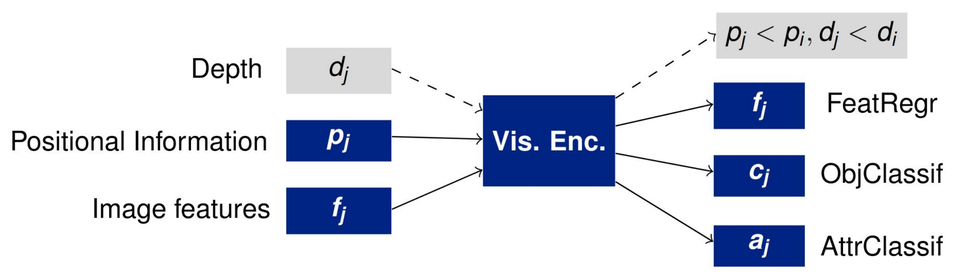
CHARLES UNIVERSITY

Positional Information in VL Models

- Missing structural analysis of Positional information (PI) of image objects in VL models: bounding box, but also width/height/area
- PI is not part of pre-training objective
- Object depth for 3D localisation is not used

- We test: no PI, center value, bounding box (bb), bb+depth
- **Probe: Original LXMERT setting**
- **Pretraining: Adding 1. and 2. as new pre-training objective**





1. Mutual Position Evaluation intrinsic, unimodal

- Nine classifications if obj_i is left (X), above (Y), behind (Z), etc. of obj_j for each object pair
- **Probe**:
  - No PI surprisingly good (80.0%) → image features as proxy
  - Center values sufficiently good (88.5%)
  - Adding depth helps for Z tasks (+4.4%P)
- **Pretraining**:
  - Better results with same pattern (88.2-93.9%)

36 objects per image

2. Contrastive Evaluation on PI using Cross-Modality Matching intrinsic, multimodal

- Permute PI word in caption with antonym
- **Probe**:
  - Good results for original task (~96 %)
  - Not able to solve multimodal PI task (<1.7 %)
- **Pretraining**:
  - Able to solve multimodal tasks (>78.1 %)

Original caption: "A student works on an academic paper at her desk, computer screen glowing in the background."

3. Downstream Task Evaluation using GQA extrinsic, multimodal

- Top 1 & 5 Accuracy and on subsets with X, Y, Z keywords
- **Probe**:
  - Best result for center values (59.4 %, +0.4 %P)
  - Adding depth helps for Z questions (+0.4 %P)
- **Pretraining:**
  - Same pattern, but slightly worse (58.8 %, Z: +0.6 %P)

X: On which **side** of the picture is the lamp?

Y: Are the windows **above** a clock?

Z: Is there a bookcase **behind** the yellow flowers?

https://www.unibw.de/vis-en/naacl2022

Published in Findings of NAACL.
Presented at NAACL 2022, Seattle, Washington.

The authors gratefully acknowledge the computing time granted by the Institute for Distributed Intelligent Systems and provided on the GPU cluster Monacum One at the Bundeswehr University Munich. The work at CUNI was funded by the Czech Science Foundation, grant no. 19-26934X.