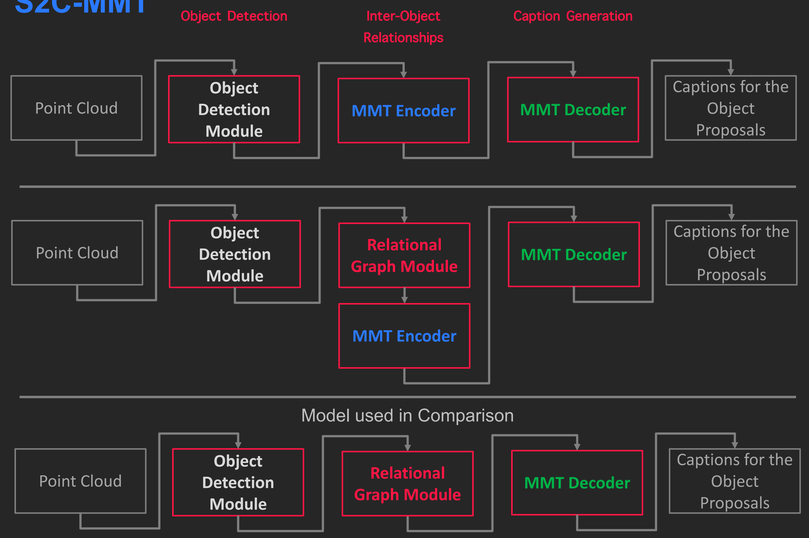
Dense Captioning for 3D Scenes with Transformers

Antonio Oroz – Kağan Küçükaytekin

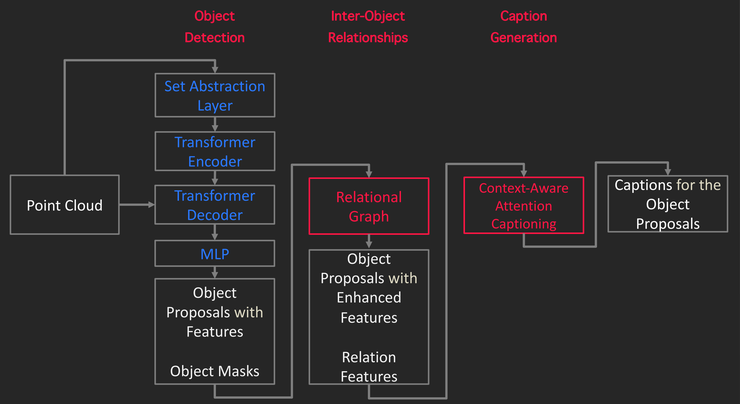
Motivation

- **S2C-MMT**
  - Transformers already proved useful in NLP tasks, for example through the ability to highly parallelizable token generation during training
  - MMT brings transformers to image captioning for 2D images
  - We want to bring it to 3D scenes
- **3DETR-S2C**
  - Recently we have seen a rise in the transformer models achieving SOTA in various computer vision tasks
  - The order independent nature of attention layers makes them suitable object detectors in point clouds.
  - We want to see if a transformer detection module can work well in a larger model

S2C-MMT

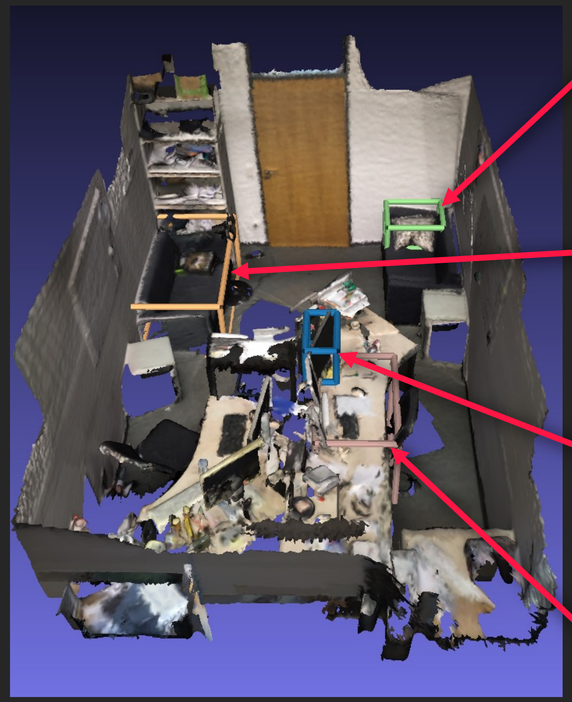


3DETR-S2C



Bounding Boxes

Examples of Generated Captions



S2C-MMT: This is a square pillow. It is on a couch.
3DETR-S2C: This is a white pillow. It is on a couch.
S2C: This is a square pillow. It is on the couch.
GT: This is a small square gray pillow. It is located on a black couch.

S2C-MMT: The couch is to the right of the door. The couch is a dark brown rectangle.
3DETR-S2C: This is a black couch. It is facing a table.
S2C: This is a brown couch. It is to the left of a brown table.
GT: It is a black sofa. It is located to the wall behind the fan.

S2C-MMT: This is a black monitor. It is on a desk.
3DETR-S2C: This is a black monitor. It is on a desk.
S2C: This is a black monitor. It is on a desk.
GT: A black computer screen is sitting on the desk. It is next to a black framed computer screen and to the left of it.

S2C-MMT: This is a brown desk. It is to the right of a chair.
3DETR-S2C: This is a black keyboard. It is on a desk.
S2C: This is a black office chair. It is in front of a desk.
GT: This is a long tan desk. It is located near a wall and a small cabinet.

Performance Comparison

	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU	mAP@0.5IoU
Scan2Cap	39.08	23.32	21.97	44.78	32.21
3DETR-S2C	41.53	**26.63**	**23.07**	**47.60**	38.61
S2C-MMT	**44.17**	24.34	22.30	45.36	**38.72**

Our two proposed models improve the performance over the baseline in all metrics.

Conclusions

- Both models offer performance improvements over Scan2Cap
- A combination of 3DETR with MMT might be promising

Future Work

- Combine 3DETR with MMT
- Test different ways to encode Inter-Object Relationships