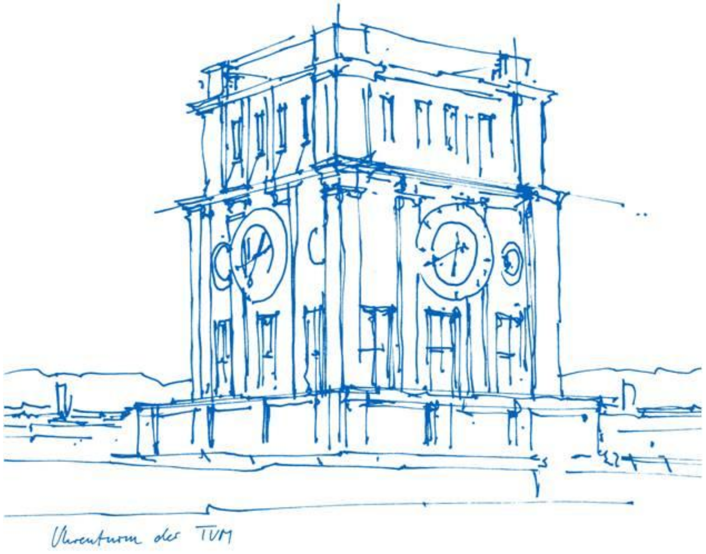
3D Visual Grounding with Transformers

1st Presentation

Advanced Deep Learning for Computer Vision (IN2364)

Department of Computer Science

Technical University of Munich



Bastian Wittmann, Philipp Foth

Munich, 31.05.21

Agenda

1. Motivation of our project
    a. Visual Grounding
    b. 3D Visual Grounding - ScanRefer
    c. 3D Object Detection - Transformer (SOTA)

2. Current Progress
    a. Roadmap
    b. Validate Claims
    c. RefNetV2
    d. Initial Results
    e. Open Challenges

Visual Grounding

Inputs:

1. Visual information (e.g. an image):

2. A natural language (NL) description:

“A man wearing a mask and and carrying a bag.”
or
“The man to the right carrying a white umbrella.”

Output: the region in the visual input corresponding to the description (e.g. a bounding box)

Visual Grounding

Task can be divided into 2 stages:

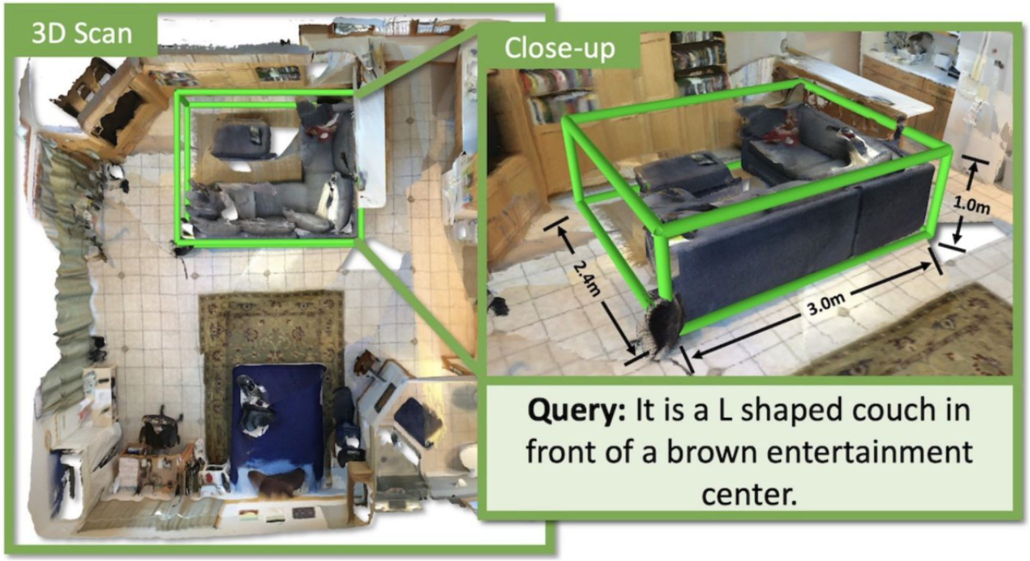
1. Object detection
2. Object localization

“A man wearing a hat and carrying a white umbrella.”

3D Visual Grounding - ScanRefer [1]

Visual input: point clouds (+ other features available)

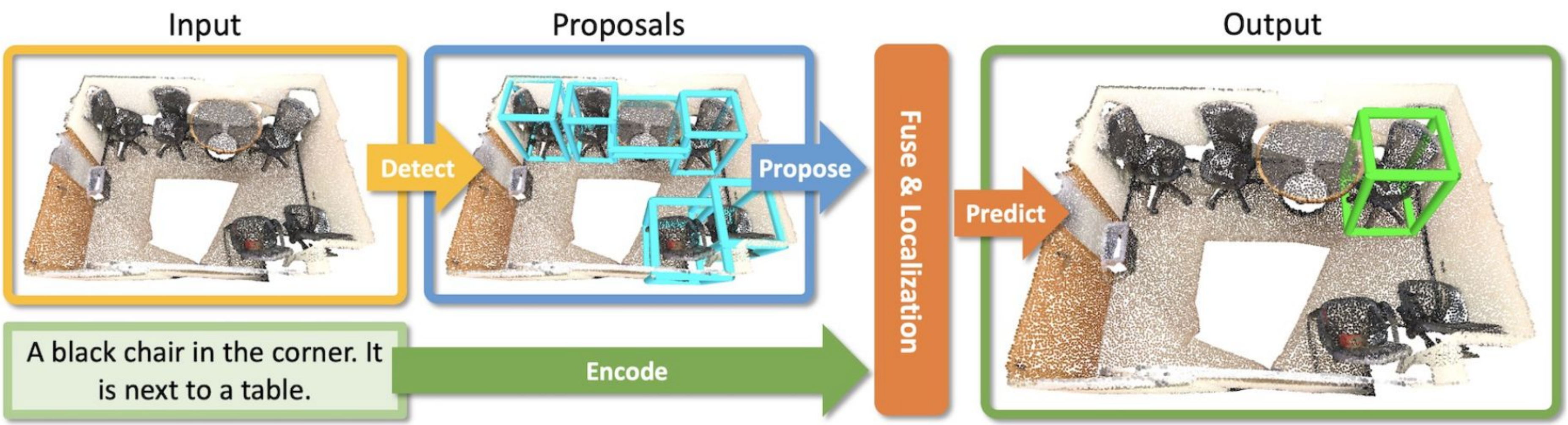
ScanRefer dataset: 51,583 descriptions of 11,046 objects from 800 ScanNet scenes



ScanRefer Method - RefNet [1]

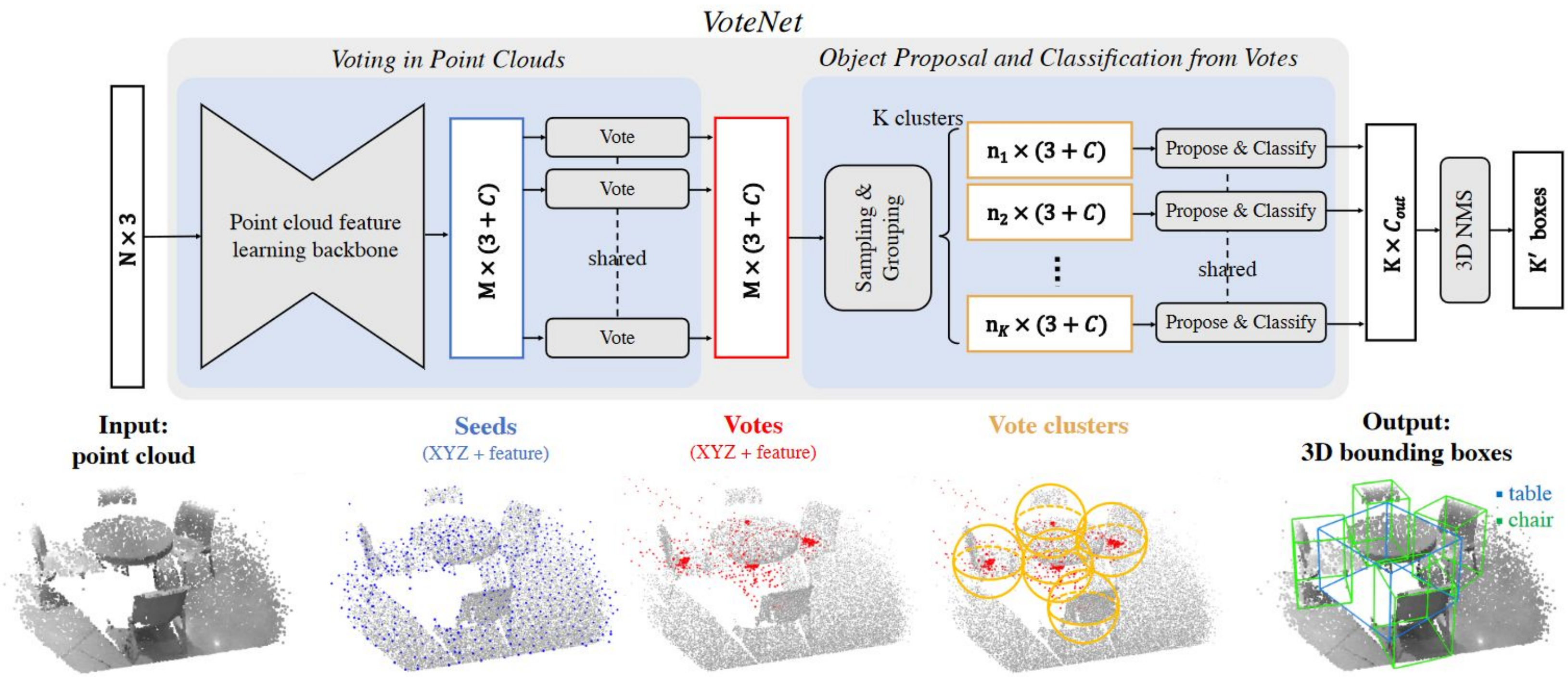
2 stages:

1. 3D object detection - **VoteNet**
2. Object localization



Achieves accuracy of 43% for IoU of 0.25

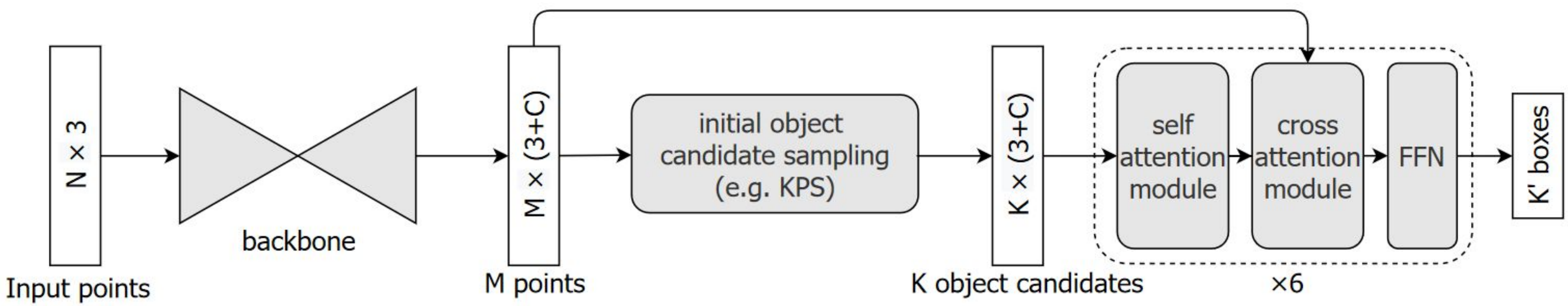
3D Object Detection - VoteNet [2]



3D Object Detection - Transformers [3]

New state-of-the-art in point cloud 3D object detection

Advantage: no “groups” are formed (group-free), rather, each object candidate can attend to all other points via the transformer



Our Tasks

Improve the **3D Visual Grounding** Performance

Bottlenecks:

- **Object detection**
  - Improve by using the SOTA detector (transformer)

- Localization
  - Design a method to incorporate language features into the transformer

Current Progress - Roadmap

Step 1

- Setup infrastructure
- Validate claims

Step 2

- Replace VoteNet
- Overfit on a few scenes
- Train using pre-trained detector
- Try to improve performance for XYZ input

Step 3

- Fine tune the pre-trained detector
- Improve performance with more input features
- Improve performance with more transformer layers and object proposals

Step 4

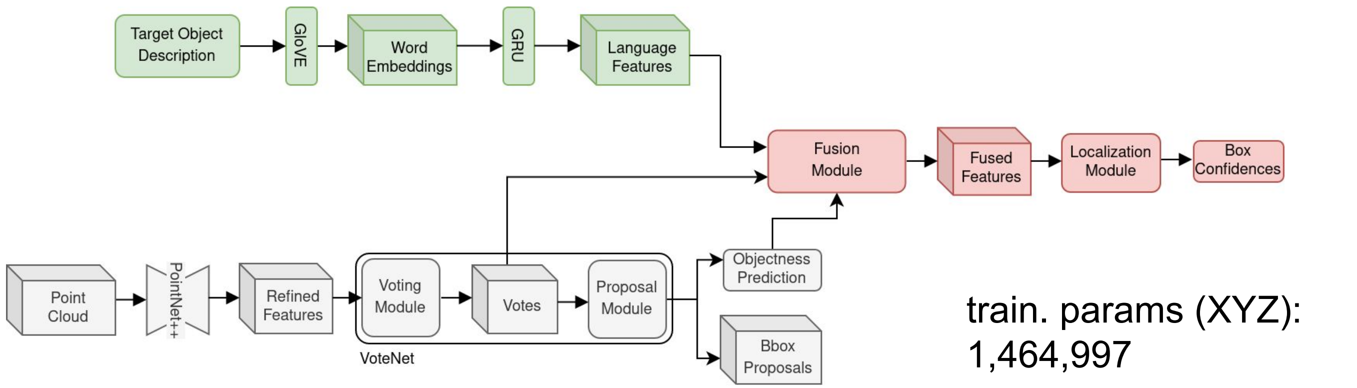
- Submit to the ScanRefer Benchmark Challenge
- Incorporate the fusion of language features with object features in the transformer

Current Progress - Validate Claims

	mAP IoU 0.25	mAP IoU 0.50	AR IoU 0.25	AR IoU 0.5	semantic cls. acc.
Transformer (XYZ) L6, O256	58.17	40.27	79.09	56.46	83.77
VoteNet (XYZ, height)	49.63	28.07	75.11	45.30	63.98
VoteNet (XYZ, height, multiview, normals)	61.89	33.53	80.67	49.13	69.84

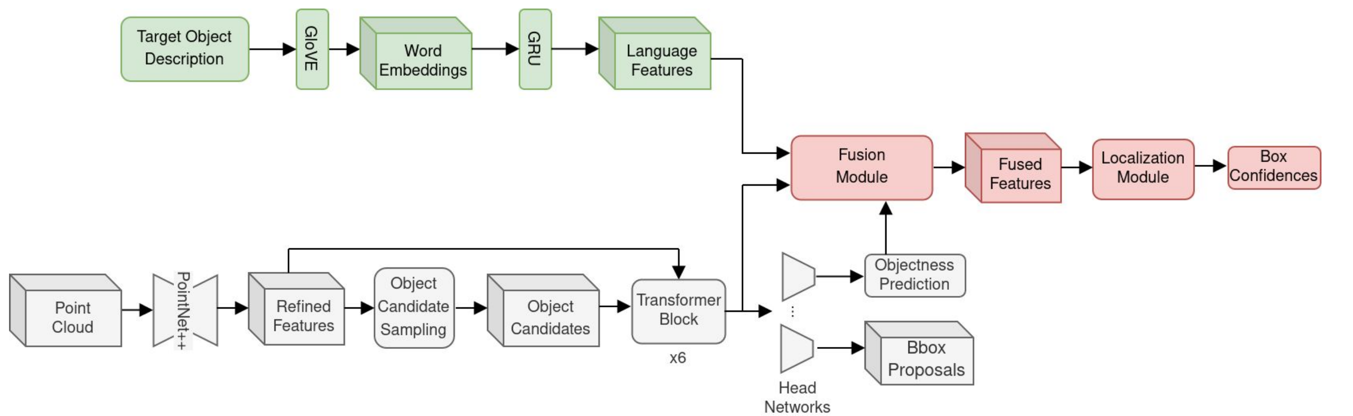
Current Progress - RefNetV2

RefNet



train. params (XYZ): 1,464,997

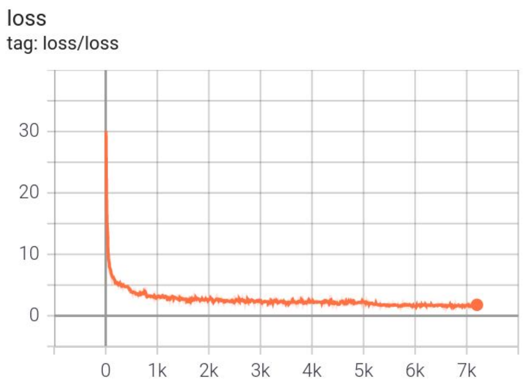
RefNetV2

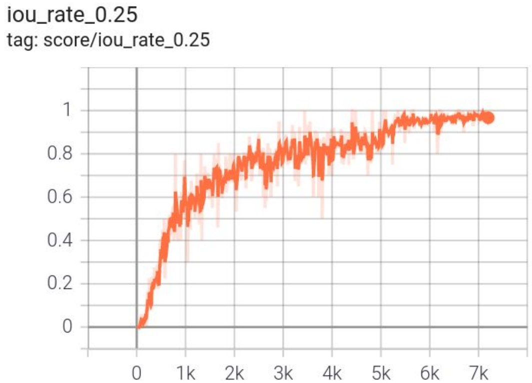


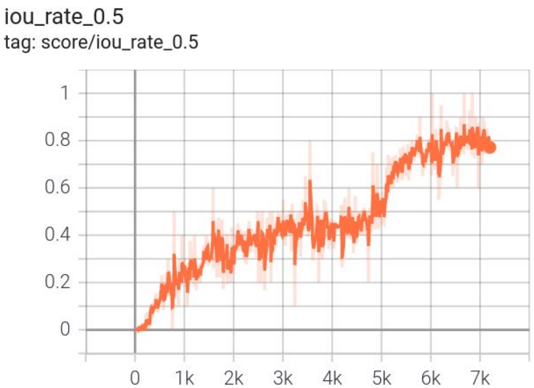
train. params (XYZ): 15,006,082

Current Progress - Initial Results

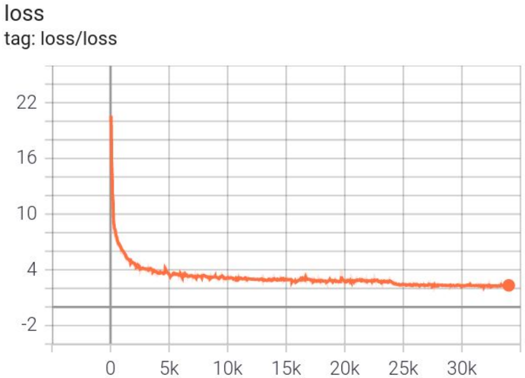
Overfit RefNetV2 to 1 scene for 400 epochs: (1 scene = multiple objects + multiple descriptions for each object)

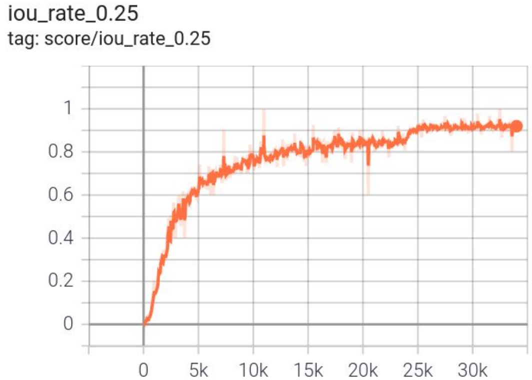


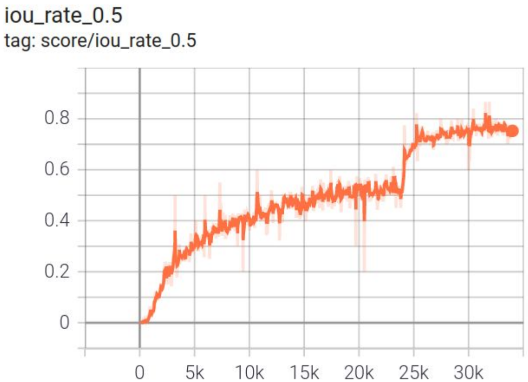




Overfit RefNetV2 to 10 scenes for 400 epochs:







Current Progress - Initial Results

	Unique Acc@0.25IoU	Unique Acc@0.50IoU	Multiple Acc@0.25IoU	Multiple Acc@0.50IoU	Overall Acc@0.25IoU	Overall Acc@0.50IoU
RefNet (XYZ, height)	63.98	43.57	29.28	18.99	36.01	23.76
RefNetV2 (XYZ) pre-trained detector L6, O256	71.04	57.12	22.22	17.35	31.70	25.06
RefNetV2 (XYZ) fine tuned detector L6, O256	*72.75* *(only pred. heads)*	*58.25* *(only pred. heads)*	*26.57* *(only pred. heads)*	*19.75* *(only pred. heads)*	*35.53* *(only pred. heads)*	*27.22* *(only pred. heads)*
RefNetV2 (XYZ) fine tuned detector L12, O512, Pointnet++ w2x	-	-	-	-	-	-
RefNet (XYZ, height, multiview, normals)	78.22	52.38	33.61	20.77	42.27	26.90
RefNetV2 (XYZ, height, rgb, normals)	-	-	-	-	-	-
RefNetV3	-	-	-	-	-	-

Current Progress - Open Challenges

Challenges:

- Pretrained transformer only takes XYZ as input
- Inconsistency between object detection evaluation of ScanRefer and the Transformer
- Inconsistency in loss functions and intermediate results
- Hard to train network end to end due the large number of trainable parameters

Next steps:

- Adapt learning rate / learning rate scheduler to current task
- Fine tune last layers of transformer to yield better results
- Leverage pretrained PointNet++ from ScanRefer to efficiently utilize additional input features

Thank you for your attention!

References

[1] Chen, D.Z., Chang, A.X., Nießner, M.: ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)

[2] Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3D object detection in point clouds. In: Proceedings of the IEEE International Conference on Computer Vision (2019)

[3] Liu, Z., Zhang, Z., Cao, Y., Hu, H., Tong, X.:Group-Free 3D Object Detection via Transformers. arXiv preprint arXiv:2104.00678 (2021)