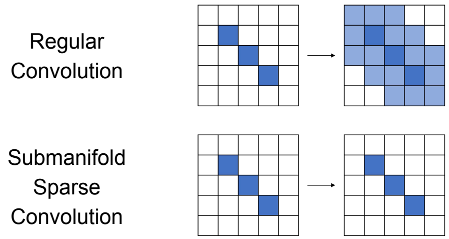
Dense Captioning on 3D Scenes with Sparse Convolutions and Reinforcement Learning

Daniel-Jordi Regenbrecht, Yifeng Dong

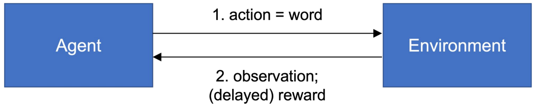
(Submanifold) Sparse Convolution

- Challenge: 3D feature extraction is difficult
- Regular convolutions too inefficient!
- But: 3D data naturally sparse → Sparse Convolution
- Submanifold Sparse Convolution preserves sparsity



Reinforcement Learning for NLP

- **Before**: Train with Cross-Entropy to predict next word given previous ground truth word
- **But**: Actually interested in NLP metrics!
- **Solution**: Use RL with CIDEr + SPICE as reward
- Loss = - Reward × log probabilities



Task: 3D Dense Captioning

- **Input:** RGB-D Point cloud.
- **Output:** List of objects + caption for each object

RL: The shelf is on the wall. The shelf is to the right of the door.

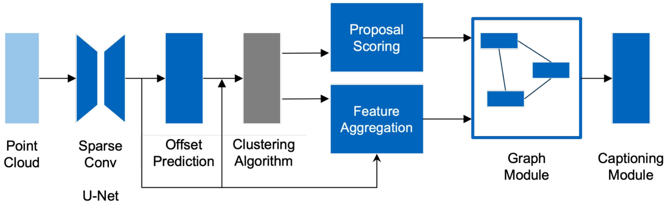
CE: This is a wooden shelf. It is above a desk.

RL: This is a brown desk. The desk is to the left of the couch.

CE: This is a brown desk. It is to the left of a chair.

Methodology

- First train **end-to-end** using typical cross-entropy loss for captions. Then **fix** pipeline and finetune only captioning module.



Results

- **Trained 3 Variants:** No RL, Warm Start, Cold Start
- Metrics: NLP Metrics thresholded by 0.5 IoU

Model	CIDEr	BLEU-4	ROUGE	Meteor
Baseline (Scan2Cap[1])	39.08	23.32	44.78	21.97
Ours (NoRL)	39.85	24.24	**46.25**	22.05
Ours (WS)	**45.76**	**26.72**	46.01	**22.51**
Ours (CS)	0.01	0.00	20.65	10.45

Conclusions

- **Submanifold Sparse Convolutions** improve object detection accuracy and overall captioning performance
- → SSCs can extract adequate features for captioning
- **Reinforcement Learning** is **feasible** for visual captioning and leads to **significantly better performance**
- It can be sufficient to **finetune** using RL

[1] *Chen et al.*, Scan2Cap