ECCV'20
ONLINE
23-28 AUGUST 2020

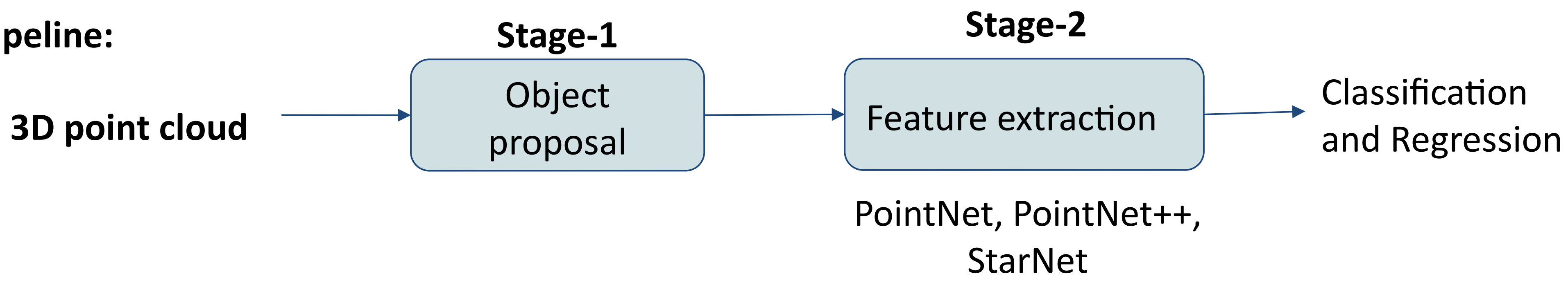
Self-Attention based Feature Extractors for 3D Object Detection in Point Clouds

Arulkumar Subramaniam[1], Ashish Vaswani[2], Niki Parmar[2]

[1]IIT Madras, India
[2]Google Brain

3D Object Detection in Point Clouds

Typical pipeline:

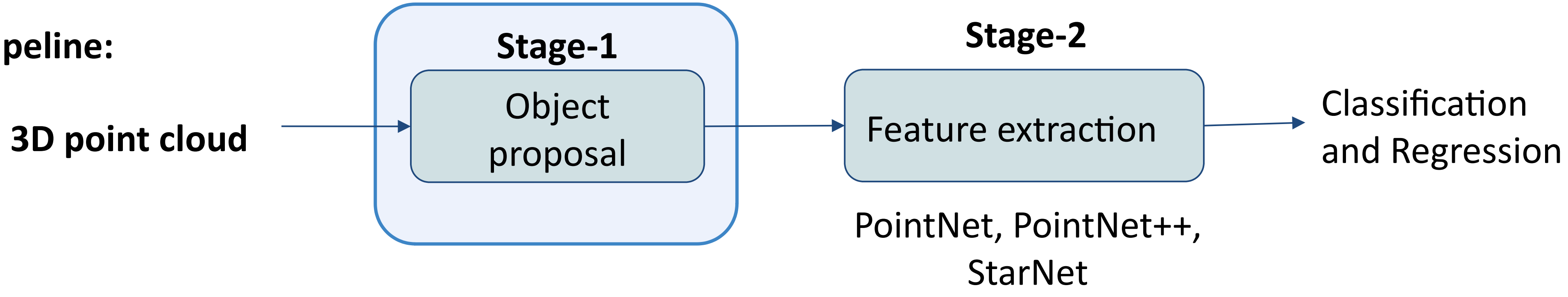


Downside:

- Point-wise feature transformations (in PointNet, PointNet++, StarNet) may not capture larger context around objects

3D Object Detection in Point Clouds

Typical pipeline:

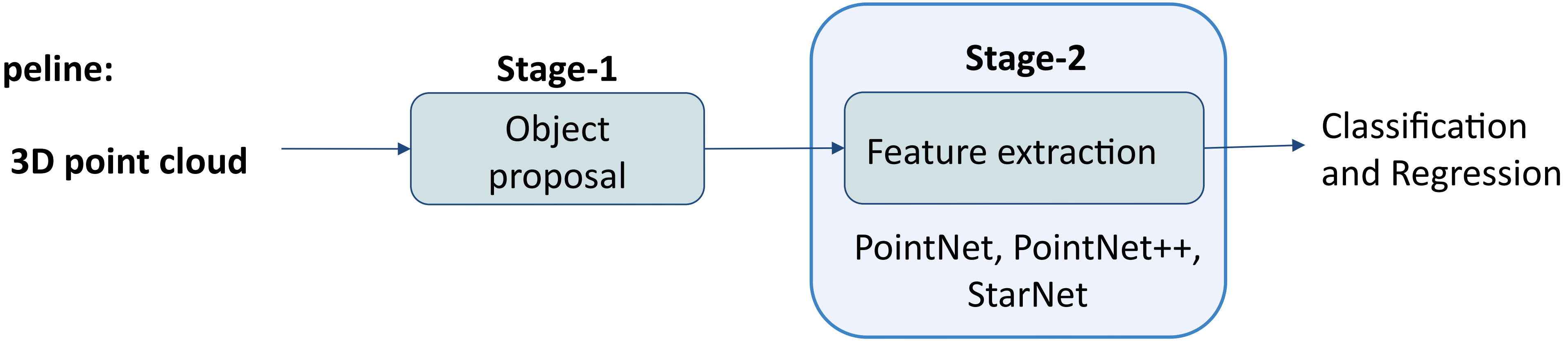


Downside:

- Point-wise feature transformations (in PointNet, PointNet++, StarNet) may not capture larger context around objects

3D Object Detection in Point Clouds

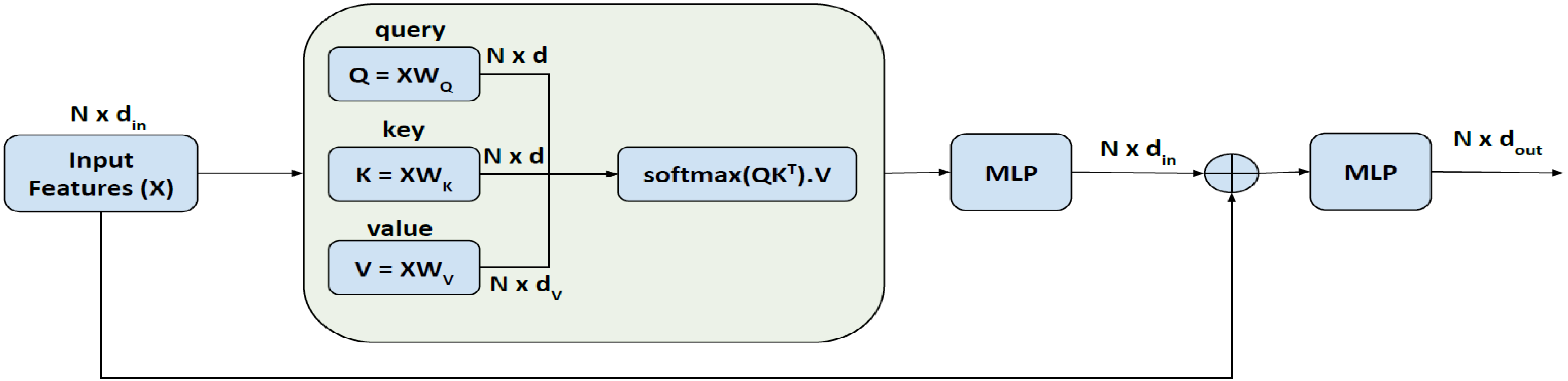
Typical pipeline:



Downside:

- Point-wise feature transformations (in PointNet, PointNet++, StarNet) may not capture larger context around objects

Self-Attention[1]

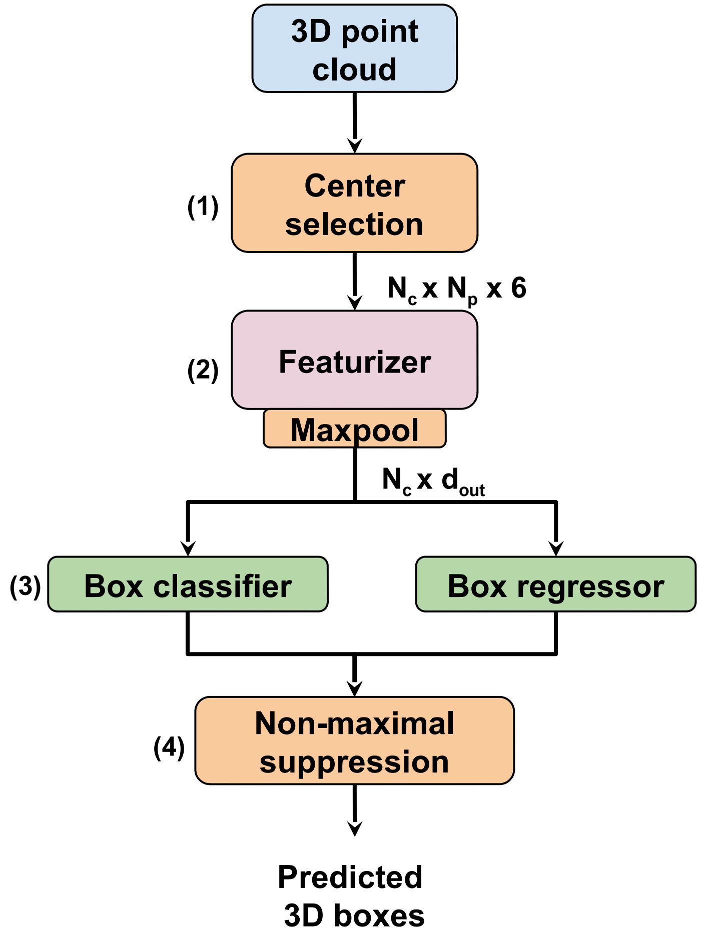


Self-attention captures local and global dependencies effectively.

In this work, we study self-attention based feature extractor for 3D object detection

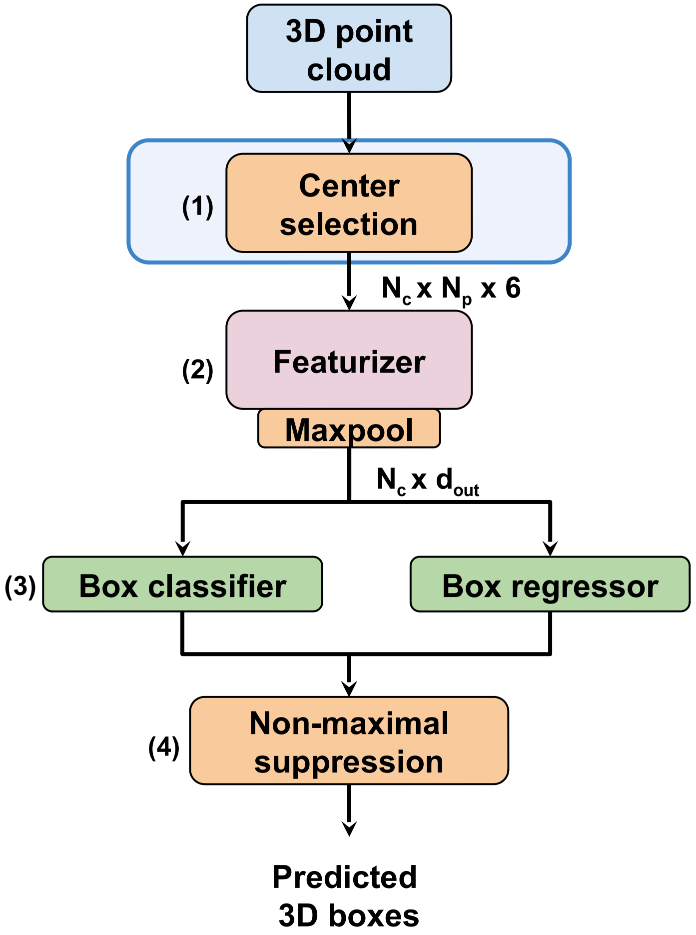
[1]Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

3D Object Detection Pipeline[1]



[1]Pipeline inspired from Ngiam, Jiquan, et al. "Starnet: Targeted computation for object detection in point clouds." *arXiv preprint arXiv:1908.11069* (2019).

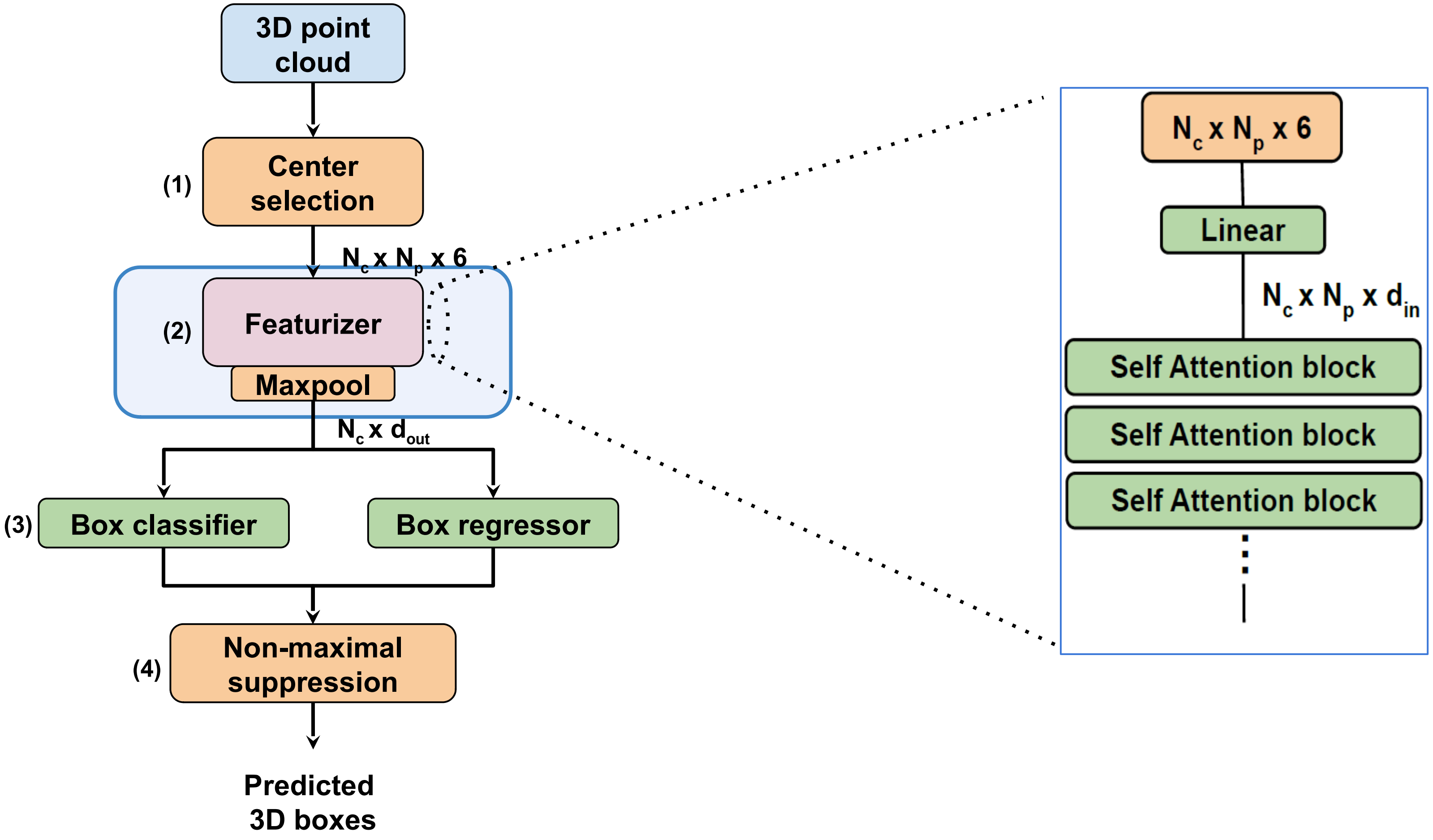
3D Object Detection Pipeline[1]



→ Farthest point sampling
N_c - Number of centers
N_p - Number of neighborhood points around center
6 - Input channels (x, y, z, range, elongation, intensity)

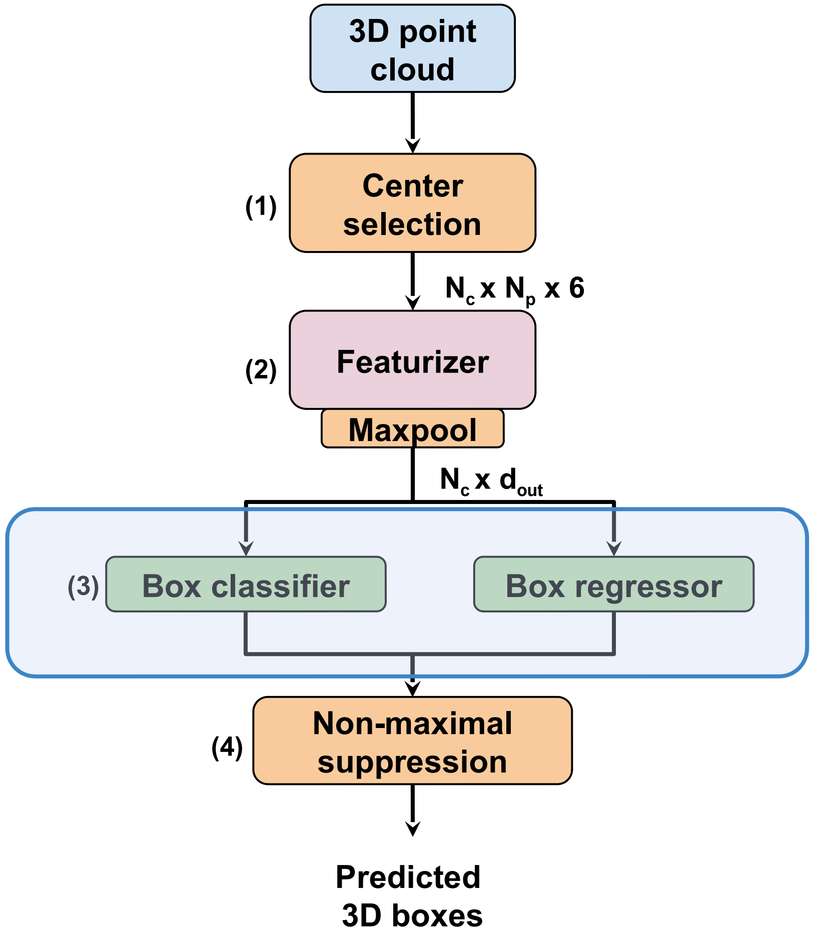
[1]Pipeline inspired from Ngiam, Jiquan, et al. "Starnet: Targeted computation for object detection in point clouds." *arXiv preprint arXiv:1908.11069* (2019).

3D Object Detection Pipeline[1]



[1]Pipeline inspired from Ngiam, Jiquan, et al. "Starnet: Targeted computation for object detection in point clouds." *arXiv preprint arXiv:1908.11069* (2019).

3D Object Detection Pipeline[1]



[1]Pipeline inspired from Ngiam, Jiquan, et al. "Starnet: Targeted computation for object detection in point clouds." *arXiv preprint arXiv:1908.11069* (2019).

Self-attention block types

1. **Neighborhood Self-attention Block (NA-block):**

   Input (X): N_c x N_p x d_{in}

   Shared self-attention block is applied on each center's neighborhood points (N_p x d_{in}) to model local dependencies (e.g., Shape)

2. **Proposal Self-attention Block (PA-block):**

   Input (X): N_c x d_{in}
   (obtained by avg-pool of neighborhood point's features N_c x N_p x $d_{in} \rightarrow N_c$ x d_{in})

   Self-attention block is applied on features of all centers to gain global context

Self-attention block types

1. **Neighborhood Self-attention Block (NA-block):**

   Input (X): $N_c \times N_p \times d_{in}$

   Shared self-attention block is applied on each center's neighborhood points ($N_p \times d_{in}$) to model local dependencies (e.g., Shape)

2. **Proposal Self-attention Block (PA-block):**

   Input (X): $N_c \times d_{in}$
   (obtained by avg-pool of neighborhood point's features $N_c \times N_p \times d_{in} \rightarrow N_c \times d_{in}$)

   Self-attention block is applied on features of all centers to gain global context

Self-attention block types

1. **Neighborhood Self-attention Block (NA-block):**

   Input (X): N_c x N_p x d_{in}

   Shared self-attention block is applied on each center's neighborhood points (N_p x d_{in}) to model local dependencies (e.g., Shape)

2. **Proposal Self-attention Block (PA-block):**

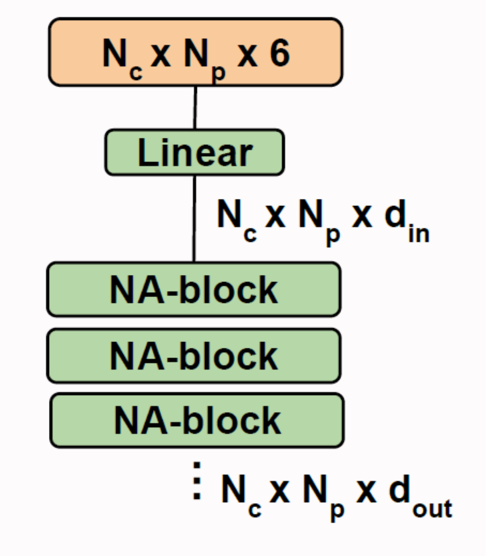
   Input (X): N_c x d_{in}
   (obtained by avg-pool of neighborhood point's features N_c x N_p x $d_{in} \rightarrow N_c$ x d_{in})

   Self-attention block is applied on features of all centers to gain global context

Proposed featurizers

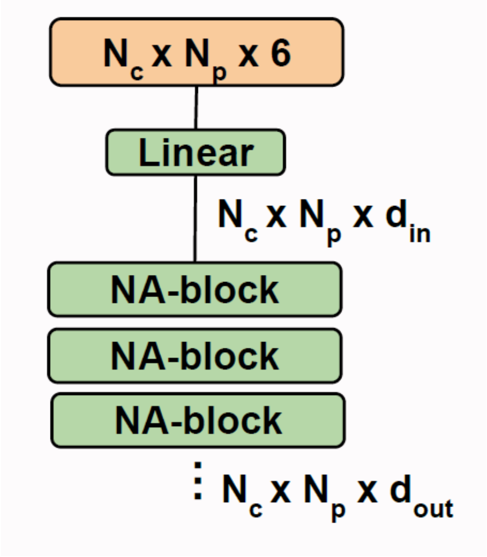
Proposed featurizers

1) **NA-only featurizer**

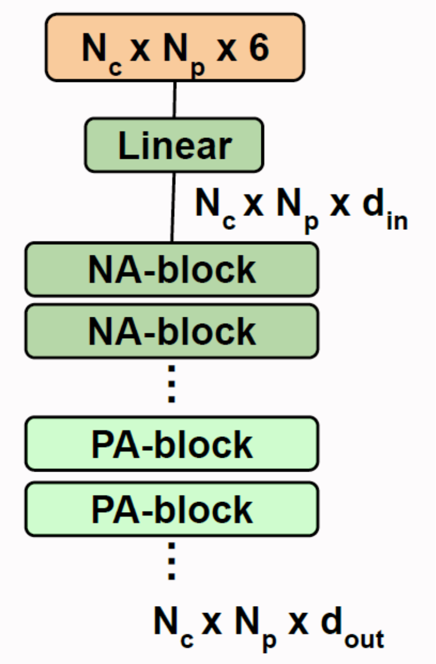


Proposed featurizers

1) **NA-only featurizer**

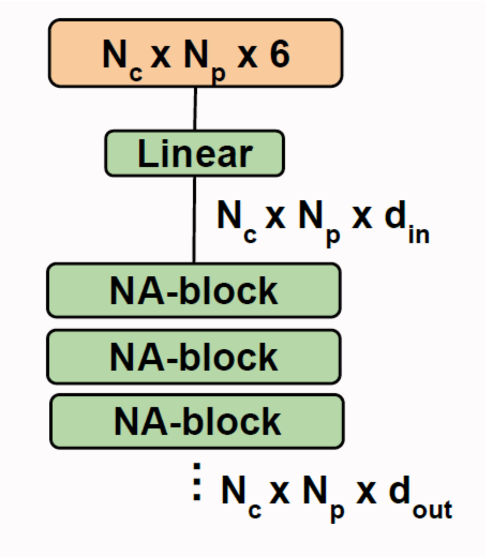


2) **NA-PA featurizer**

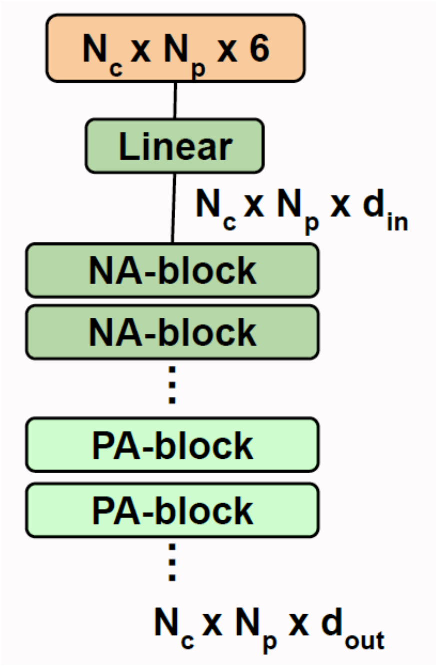


Proposed featurizers

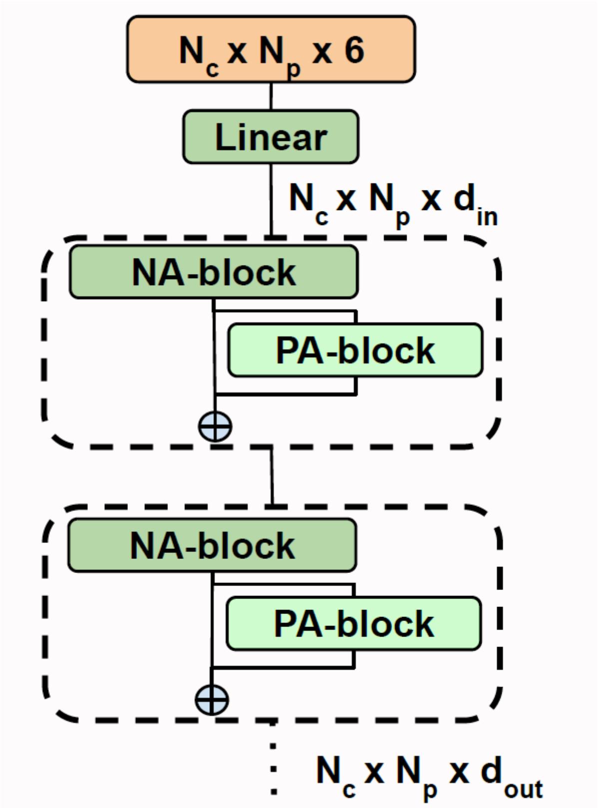
1) NA-only featurizer



2) NA-PA featurizer



3) NA-PA Alternated featurizer



Dataset, Evaluation protocol

Waymo Open Dataset:

- 1000 segments of 20 seconds LiDAR measurements (rate of 10 Hz)
- In our experiments, we use two classes:
    - **Pedestrian**
    - **Vehicle**
- **Evaluation metric:** mean average precision (mAP) on Waymo validation set
- The hyperparameters are same as StarNet (Ngiam et. al)
  For our models, Number of attention heads = 4
- For testing, we use N_c = 1024, N_P = 256

Comparison with state-of-the-art methods

Models	#params	#GFlops	Pedestrian mAP	Vehicle mAP
PointPillars[1]	-	3700	62.1	57.2
MVF[2]	-	-	65.3	**62.9**
StarNet[3]	1.483 M	136.56	66.8	53.7
NA-only featurizer ($N_{NA} = 4$)	0.316 M	128.7	67.83	53.91
NA-only featurizer ($N_{NA} = 10$)	0.467 M	317.15	**69.05**	59.2
NA-PA featurizer ($N_{NA} = 4, N_{PA} = 4$)	0.421 M	130.08	67.64	58.09
NA-PA Alternated featurizer ($N_{alternate} = 4$)	0.421 M	131.8	68.3	58.66

Table 1: Comparisons on Waymo validation set. $N_{NA}, N_{PA}, N_{alternate}$ are number of NA-, PA-, Alternated NA and PA blocks respectively.

Comparison with state-of-the-art methods

Models	#params	#GFlops	Pedestrian mAP	Vehicle mAP
PointPillars[1]	-	3700	62.1	57.2
MVF[2]	-	-	65.3	**62.9**
StarNet[3]	1.483 M	136.56	66.8	53.7
NA-only featurizer ($N_{NA} = 4$)	0.316 M	128.7	67.83	53.91
NA-only featurizer ($N_{NA} = 10$)	0.467 M	317.15	**69.05**	59.2
NA-PA featurizer ($N_{NA} = 4, N_{PA} = 4$)	0.421 M	130.08	67.64	58.09
NA-PA Alternated featurizer ($N_{alternate} = 4$)	0.421 M	131.8	68.3	58.66

Table 1: Comparisons on Waymo validation set. $N_{NA}, N_{PA}, N_{alternate}$ are number of NA-, PA-, Alternated NA and PA blocks respectively.

Comparison with state-of-the-art methods

Models	#params	#GFlops	Pedestrian mAP	Vehicle mAP
PointPillars[1]	-	3700	62.1	57.2
MVF[2]	-	-	65.3	**62.9**
StarNet[3]	1.483 M	136.56	66.8	53.7
NA-only featurizer ($N_{NA} = 4$)	0.316 M	128.7	67.83	53.91
NA-only featurizer ($N_{NA} = 10$)	0.467 M	317.15	**69.05**	59.2
NA-PA featurizer ($N_{NA} = 4, N_{PA} = 4$)	0.421 M	130.08	67.64	58.09
NA-PA Alternated featurizer ($N_{alternate} = 4$)	0.421 M	131.8	68.3	58.66

Table 1: Comparisons on Waymo validation set. $N_{NA}, N_{PA}, N_{alternate}$ are number of NA-, PA-, Alternated NA and PA blocks respectively.

Comparison with state-of-the-art methods

Models	#params	#GFlops	Pedestrian mAP	Vehicle mAP
PointPillars[1]	-	3700	62.1	57.2
MVF[2]	-	-	65.3	**62.9**
StarNet[3]	1.483 M	136.56	66.8	53.7
NA-only featurizer ($N_{NA} = 4$)	0.316 M	128.7	67.83	53.91
NA-only featurizer ($N_{NA} = 10$)	0.467 M	317.15	**69.05**	59.2
NA-PA featurizer ($N_{NA} = 4, N_{PA} = 4$)	0.421 M	130.08	67.64	58.09
NA-PA Alternated featurizer ($N_{alternate} = 4$)	0.421 M	131.8	68.3	58.66

Table 1: Comparisons on Waymo validation set. $N_{NA}, N_{PA}, N_{alternate}$ are number of NA-, PA-, Alternated NA and PA blocks respectively.

Thank you!

Self-Attention based Feature Extractors for 3D Object Detection in Point Clouds

Arulkumar Subramaniam, Ashish Vaswani, Niki Parmar