

Towards Efficient Point Cloud Graph Neural Networks Through Architectural Simplification

Shyam A. Tailor*
 University of Cambridge
 sat62@cam.ac.uk

René de Jong Tiago Azevedo Matthew Mattina Partha Maji
 Arm ML Research Lab
 first.last@arm.com

Abstract

In recent years graph neural network (GNN)-based approaches have become a popular strategy for processing point cloud data, regularly achieving state-of-the-art performance on a variety of tasks. To date, the research community has primarily focused on improving model expressiveness, with secondary thought given to how to design models that can run efficiently on resource constrained mobile devices including smartphones or mixed reality headsets. In this work we make a step towards improving the efficiency of these models by making the observation that these GNN models are heavily limited by the representational power of their first, feature extracting, layer. We find that it is possible to radically simplify these models so long as the feature extraction layer is retained with minimal degradation to model performance; further, we discover that it is possible to improve performance overall on ModelNet40 and S3DIS by improving the design of the feature extractor. Our approach reduces memory consumption by $20\times$ and latency by up to $9.9\times$ for graph layers in models such as DGCNN; overall, we achieve speed-ups of up to $4.5\times$ and peak memory reductions of 72.5%.

1. Introduction

3D scanning technology is rapidly becoming ubiquitous as the underlying sensors become smaller, cheaper, and more efficient [2]. These sensors can produce point clouds, which are unordered sets of points in 3D space. This is in contrast to other common modalities, such as images, where we have a regular grid structure. Handling this unstructured representation of data is difficult, and a variety of approaches have been proposed. Early approaches in the literature include projecting the points onto planes, and applying convolutional neural networks (CNNs) [26], or voxelizing the point cloud and applying 3D convolutions [23].

In recent years, however, a dominant approach has been to construct graphs from the point cloud, and apply graph neural networks (GNNs) [21, 33]. These approaches continue to achieve high performance on tasks including object classification, part segmentation, and semantic segmentation, with many recent works investigating variations in layer design to improve model accuracy [38, 40].

Given the rising popularity of 3D scanning sensors in mobile devices such as smartphones [1] or mixed reality headsets [24], it is natural to investigate how to design resource-efficient models that can run *on-device*. Unlike some other fields of computer vision, there has been relatively little research into how to design more efficient models, with the majority of works exclusively focusing on how to increase model accuracy, with secondary thought given to considerations such as memory consumption or latency. This motivates our work: we seek to understand where it is possible to apply simplifications to point cloud architectures. With a greater understanding of which design choices affect the accuracy of the models, we aim to radically simplify popular existing model architectures, as a first step towards enabling on-device models for point cloud data.

A key challenge with GNNs is their resource demands. Many state-of-the-art GNNs [32, 7], including point-cloud specific models [33, 21] require memory and OPs proportional to the number of edges in the graph ($\mathcal{O}(E)$), as they use complicated anisotropic mechanisms to create the messages that nodes send to each other. In contrast, many simpler GNN architectures [15, 37] can be implemented using memory and OPs proportional to the number of vertices ($\mathcal{O}(V)$), and it was recently shown by Tailor et al. [29] that $\mathcal{O}(V)$ GNN architectures could surpass state-of-the-art models on a variety of tasks. Given that the number of edges is typically $20\text{--}40\times$ larger than the number of vertices for graphs in point cloud models, this represents a significant opportunity for optimization: the MLP layers constitute most of the latency [5]. This can be seen in Figure 1. Unfortunately, however, naively applying standard GNN layers to point cloud data yields poor results as they fail to incorporate geometric priors. In this work, we investigate

*Work completed during Shyam’s internship at Arm ML Research Lab.
 Correspondence to Shyam Tailor and Partha Maji.

which components are most critical to point cloud model performance so that we can incorporate optimizations already known in the wider GNN community.

Succinctly, we hypothesise that the most critical layer to model performance is the very first one acting on the raw point cloud data. We find support for our hypothesis by evaluating it on two datasets, ModelNet40 and S3DIS, with two models, PointNet++ [21] and DGCNN [33]. Although this hypothesis is relatively simple, it has two useful implications: firstly, if we aim to increase accuracy, it is worth investing more resources into this feature extracting block, and specialising it for the input data. Secondly, it is possible to simplify the rest of the layers in the model, with little impact on accuracy, but offering large improvements in efficiency. Together, these observations provide a strong step towards designing more efficient GNN-based models for point cloud data.

In summary, our work makes the following contributions:

1. We provide extensive ablation studies on two of the most popular GNN-based architectures to point cloud processing, PointNet++ and DGCNN. Our experiments demonstrate the validity of our hypothesis regarding the importance of the feature extraction layer.
2. We demonstrate how to improve the feature extraction layer by incorporating stronger geometric priors. Our approach enables DGCNN to achieve close to state-of-the-art results on ModelNet40, and yields improvements of more than 2 mIoU on S3DIS.
3. We illustrate that our approach reduces memory consumption by up to 72.5% in practice. We achieve speed-ups of up to $4.2\times$ on CPU, and $2.5\times$ on GPU.

2. Related Work

2.1. Architectures for Point Clouds

Multiple approaches have been proposed for processing point cloud data using neural networks; this section describes the most popular approaches.

Mapping 3D Points to 2D Images Multi-view CNN [26] approaches involve projecting the point cloud onto 2D planes, and feeding the resulting depth images into a CNN model. In recent years, these approaches have fallen out of favour, however a recent work has shown that they still offer a strong baseline for point cloud classification [10]. Despite this, there are significant weaknesses associated with these methods: they struggle with non-uniform or low point density, and occlusions.

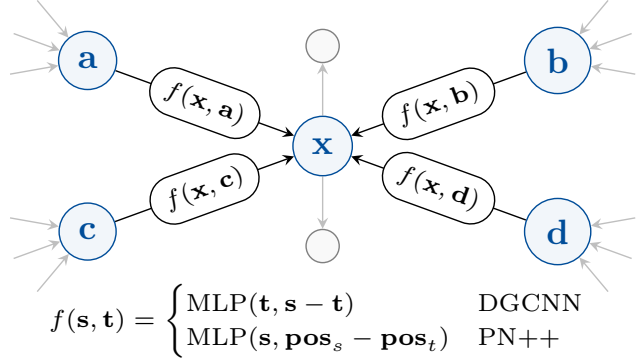


Figure 1. Popular approaches for handling point clouds with GNNs require each message to be explicitly computed and materialized as they are a function of both the source (\mathbf{s}) and target (\mathbf{t}) nodes. This significantly increases memory and latency. We illustrate in this paper that many of these operations can be safely removed so long as an effective feature extraction layer is retained.

3D Convolutions Another approach is to discretize the 3D space into voxels, and apply 3D convolutions [23, 19, 4, 12]. The primary weakness with this approach is the inevitable information loss when voxelizing the point cloud; this can be mitigated by using a finer discretization, but this results in greater memory and computational costs.

Point Convolutions PointCNN [16] proposed learning transformations that enable the convolution to be done in a different, regular, space; however, this approach does not retain permutation-invariance, which restricts representational power. Other approaches [39, 34] define *continuous* kernels; implementation requires explicit prediction of the kernel weights, resulting in high computational cost. KP-Conv [30] proposes to simplify these approaches by restricting the convolution to be defined by a set of anchor points, reducing the computational complexity and improving accuracy.

Graph Neural Networks PointNet [20] proposed directly operating on the input points using shared MLPs, and using global pooling operations to share information across the point cloud. This approach is highly efficient, but does not allow for local information to be aggregated, restricting the expressiveness of the architecture. Follow-up works, such as PointNet++ [21] and DGCNN [33] built upon PointNet by explicitly constructing graphs upon which local aggregation is performed, thereby boosting accuracy at the cost of significant increases to memory consumption and runtime. These models can be interpreted using the message passing paradigm that is commonly used for describing graph neural networks (GNNs) [9]. In this paradigm, a source node sends a *message* along a directed edge to a target node; each node *aggregates* the messages it has re-

ceived using a permutation-invariant function, such as max or mean, and uses the aggregated result to update its representation. For PointNet++ and DGCNN we have the following rules for updating node representations:

$$\mathbf{x}_i^{l+1} = \max_{j \in \mathcal{N}(i)} \text{MLP}(\mathbf{x}_j^l \parallel \mathbf{p}_j - \mathbf{p}_i) \quad (\text{PointNet++})$$

$$\mathbf{x}_i^{l+1} = \max_{j \in \mathcal{N}(i)} \text{MLP}(\mathbf{x}_i^l \parallel \mathbf{x}_j^l - \mathbf{x}_i^l), \quad (\text{DGCNN})$$

where \mathbf{x}_i^l is the representation of point i at layer l , \mathbf{p}_i represents the 3D position of point i , and $\mathcal{N}(i)$ is the set of neighbors of point i in the constructed graph, which is found using kNN for DGCNN and radius queries for PointNet++. In the first layer, DGCNN represents \mathbf{x}_i as the point features (if any) concatenated with the point position.

In this work we focus on PointNet++ and DGCNN, which are among the most popular architectures, and remain strong baselines in many cases [10]. However, the approaches and observations described in this work are agnostic to architecture, and are transferable to more recent approaches achieving state-of-the-art performance [38, 40]. It is worth noting that our conclusions will also be useful to point convolution approaches, which also operate on graphs and can be cast as a sub-group of the GNN approaches.

2.2. Optimizing Graph Neural Networks

There is relatively little work on optimizing GNN inference. Quantization [27] is one approach to reduce latency and memory consumption by reducing the bitwidth of the weights and activations. In practice this is not straightforward for GNNs, with specialist training procedures being required [28]. Another avenue of research is architecture design: Zhao et al. [41] studied how to arrange GNN layers to maximise accuracy given certain constraints. Taylor et al. [29] explicitly studied efficient architecture design for GNNs, and proposed an architecture that could achieve state-of-the-art performance while using memory and OPs proportional to the number of vertices in the graph—not the number of edges as is common for many GNN architectures [35], including PointNet++ and DGCNN.

Point Cloud Model Optimization Li et al. [17] demonstrated that it is possible to significantly reduce the OPs used by DGCNN by decomposing the MLP operation into two separate MLPs: $\text{MLP}_1(\mathbf{x}_i^l) - \text{MLP}_2(\mathbf{x}_j^l)$. This observation is effective at reducing OPs, but does not reduce memory consumption at inference time, and this work did not investigate how this approach generalized to other architectures. Another recent work has investigated accelerating the grouping operations these architectures use to construct the graph: KeOps [5] achieves a 10-100 \times speed-up in practice

for these operations. In practice, the latency bottleneck in these models is now the MLP layers and aggregation operations [5], which are the focus of this work. Accelerating the MLP layers and aggregations is possible with quantization, which has been studied for point cloud networks by [22]. PosPool [18] demonstrated that it is possible to replace the sophisticated aggregation methods proposed by many point cloud works with a simpler method; however, to achieve this, they relied upon a deep residual network, which has significantly higher latency due to the number of aggregation steps. Finally, RandLa-Net [14] demonstrated that by encoding individual points in the point cloud with details of their neighborhood, it was possible to use a random down-sampling, rather than the slower (but more uniform) farthest point sampling technique, to achieve high accuracy when segmenting large point clouds.

3. Why Are Point Cloud GNNs Intolerant of Optimizations?

Although works such as EGC [29] have shown great promise at optimizing GNNs for non-point cloud tasks, our early experiments indicated that they do not achieve strong performance on point cloud tasks. Indeed, we have observed a proliferation of architectures designed for point clouds exclusively, a trend that has not occurred to nearly the same extent for other data modalities GNNs have been applied to. We ask: **why do we need point cloud-specific architectures?**

3.1. Our Hypothesis: Information Bottleneck at the First Layer

We hypothesise that point cloud architectures are severely bottlenecked by the information extracted at the first layer, where the primary input is the point positions. Following this hypothesis, we believe that the reason point cloud-specific architectures achieve better results is because they incorporate geometric priors into this first layer. Critically, we argue that these geometric priors lose relevance deeper into the network when reasoning about the representations geometrically is less well defined—and are arguably even harmful given their lack of success on other data modalities, and potential to induce overfitting. There are two important corollaries of this hypothesis:

1. To improve accuracy it is best to spend resources improving the feature extracting first layer.
2. After the first layer, it is possible to radically simplify the architecture to reduce latency and memory consumption.

3.2. Experimental Verification

We now provide experimental evidence supporting our hypothesis by evaluating DGCNN and PointNet++ on Mod-

Model	Experiment	ModelNet40		S3DIS Area 5		
		OA	mAcc	mIoU	OA	mAcc
PN++ SSG	Baseline	92.8	89.5	55.2	84.6	64.0
PN++ MSG	Baseline	92.5	89.4	-	-	-
DGCNN	Baseline	92.9	88.9	53.1	85.3	59.8
PN++ SSG	(1)	90.5	86.8	54.0	84.1	62.2
PN++ MSG	(1)	91.3	87.8	-	-	-
DGCNN	(1)	91.7	88.9	49.2	83.3	57.6
PN++ SSG	(2)	91.2	88.7	54.3	84.6	61.8
PN++ MSG	(2)	90.9	87.7	-	-	-
DGCNN	(2)	92.8	89.3	51.6	83.9	57.6
PN++ SSG	(3)	92.9	89.3	54.3	84.5	62.3
PN++ MSG	(3)	92.6	88.6	-	-	-
DGCNN	(3)	92.8	90.4	54.0	85.2	62.5

Table 1. Results of inserting simplified blocks into PointNet++ and DGCNN. SSG and MSG refer to single and multi-scale grouping PointNet++ variants respectively [21]. Experiment (1) refers to completely replacing all blocks with simplified versions; (2) refers to replacing only the first block; (3) refers to replacing all blocks *except the first one*. Retaining the first block only yields results comparable, if not better, than the baseline.

elNet40 [36] and S3DIS Area 5 [3]. As pointed out by Goyal et al. [10], precise experimental setup can have a large effect on the end results obtained when working with point cloud data. We use a similar protocol to [38]; one change we make is to sample points for ModelNet40 randomly, rather than using farthest point sampling, to make the task more difficult. To facilitate the interpretation of the results, we run the baselines with our implementations to ensure our results are self-consistent.

Many works in the point cloud literature have argued that “centralization” operations, in which node representations are centralized to their local neighborhood, are vital to obtaining strong performance [21, 33, 18]; these can be witnessed in the update rules for both PointNet++ and DGCNN. However, performing these centralization operations has significant memory and computational overheads, as seen in fig. 1. Therefore, we devote this section to assessing when exactly centralization is required, enabling us to evaluate our hypothesis. In particular, we consider simplified update rules for both PointNet++ and DGCNN:

$$\mathbf{x}_i^{l+1} = \max_{j \in \mathcal{N}(i)} \text{MLP}(\mathbf{x}_j^l \parallel \mathbf{p}_j) \quad (\text{PointNet++-Simple})$$

$$\mathbf{x}_i^{l+1} = \max_{j \in \mathcal{N}(i)} \text{MLP}(\mathbf{x}_j^l) \quad (\text{DGCNN-Simple})$$

The reader should also note that both these simplified operators significantly reduce the memory consumption and OP count. These simplified operators can be implemented using $\mathcal{O}(V)$ memory and OPs, and the propagation step can be implemented using matrix multiplication-style algorithms, which are more practical to accelerate [29]. In con-

trast, the baseline operators use $\mathcal{O}(E)$ memory and OPs. Therefore, if our hypothesis holds, we can achieve savings of approximately 20-40 \times in both quantities compared to the baselines in the graph layers.

We present experiments on ModelNet40 and S3DIS Area 5 in Table 1. We focus on ablating the contribution of the baseline geometrically-inspired update rules, by evaluating their contribution at the first layer, and beyond the first layer. Our experiments clearly demonstrate that using a simplified block after the first layer does not significantly reduce accuracy, and in some cases, actually improves it. In contrast, when a simplified block is used as the first layer, there is a noticeable drop in accuracy. It is also not the case that a sophisticated backbone can recover the information lost in the first layer: we observe only changing the first layer still results in a noticeable accuracy drop. This trend is especially clear for DGCNN; unlike PointNet++, there is no additional geometric information injected at later layers of the model, in the form of point positions, hence we expect that it is more reliant on the feature extractor. We expect the performance of the simplified architectures—although superior to the baselines in many cases—can be improved further with better hyperparameters, which have been tuned for the baselines exclusively.

These results support our hypothesis, and motivate the rest of this work with a natural follow-up question: **can we improve the feature extractor to improve accuracy, while retaining low overall resource usage?**

4. Designing Accurate and Efficient Feature Extractors

In this section we investigate how to construct better feature extractors from our point cloud, and show that incorporating basic geometric information into the first feature extracting layer, and retaining a simplified backbone, can enable us to improve model accuracy. We use DGCNN as a motivating case study, however our experiments can equally be applied to other backbones; we find that our modifications enable us to obtain close to state-of-the-art performance on ModelNet40¹, and increases of more than 2 mIoU on S3DIS relative to an unmodified DGCNN model. This confirms corollary (1) from our hypothesis described in section 3.1, where we stated that improving feature extractors can enable us to achieve better accuracy overall.

4.1. Finding Useful Geometric Features

It is well known that adding information such as normal vectors or classical point cloud feature histograms to the input point cloud will boost accuracy; we are making a

¹The reader should note that our experiments are run using non-uniform point densities, making our task more difficult than most results in the literature

Source Pos	Target Pos	Rel Pos	Distance	OA	mAcc
✓	✓	✓	✓	93.2	90.2
				93.4	90.4
✓		✓	✓	93.4	90.7
✓	✓		✓	92.8	90.0
✓	✓	✓		92.9	89.9
	✓	✓		92.8	90.4

Table 2. Results of ablation study on feature extractor on ModelNet40. We assess removing each individual component of the feature extractor to assess their contribution to end performance. The bottom row is DGCNN’s standard block, which can be interpreted as implementing a subset of the operations included in our feature extractor. We observe that adding features such as distance can noticeably increase accuracy to close to state-of-the-art.

related observation that simply incorporating stronger geometric priors into the first, feature-extracting, layer of the architecture can serve as a cheap and effective way to boost accuracy [20, 25]. These geometric priors do not need to be repeated throughout the model, affording us the use of efficient graph layers elsewhere.

To investigate the contribution of different geometric priors at the first layer, we consider the use of a new update rule consisting of 4 features applied to ModelNet40. The geometric features we identify are: (1) Source position \mathbf{p}_j ; (2) Target position \mathbf{p}_i ; (3) Relative position $\mathbf{p}_j - \mathbf{p}_i$, and (4) Euclidean distance $\text{dist}(\mathbf{p}_j, \mathbf{p}_i)$. We form messages from the source to the target node by concatenating the chosen features together and applying an MLP; the reader should note that this update rule simply extends the default DGCNN update rule at the first layer with additional features. All layers other than the first use the simplified DGCNN layers described in the previous section.

The results of our ablation study on the proposed feature extractor are shown in Table 2. We observe in the top row that including all 4 features yields an increase in overall accuracy; however, removing either source or target positions from the update rule can boost accuracy further. In contrast, removing relative positional information or distance information noticeably reduces the accuracy. We believe that the increase in accuracy due to removing source or target position is due to reduced overfitting; the reader should note that providing all three positional components is redundant, as the third component can be directly computed from the other two provided. However, given that the removal of relative positional information noticeably impacts accuracy, we believe that faithful reconstruction of relative position from source and target positions cannot be relied upon without careful or lucky weight initialization. As the relative position is more critical to model performance than either the raw source or target positions, it is better to explicitly compute it before feeding into the model. Similarly,

Aggregators	mIoU	OA	mAcc
Baseline (tab. 1)	53.2	85.3	62.5
max	55.2	85.4	63.2
max + min	53.6	86.0	60.8
max + mean	55.3	84.9	63.2

Table 3. Performance of the proposed feature extractor on S3DIS. We observe that our feature extractor improves performance over the baseline DGCNN model by 2 mIoU. Adding additional aggregators, although effective on other GNN datasets [7], does not yield immediate improvements for point cloud data.

distance cannot be computed without multiple layers (as it is a non-linear transform), increasing the required OPs and memory, and making training more difficult [18].

Improving Performance on S3DIS With our improved understanding of how to handle positional data, we proceed to demonstrate that these gains translate to the more difficult S3DIS dataset. Unlike ModelNet40, where the input data consists exclusively of raw point positions, S3DIS introduces RGB color information for each point. Building upon our conclusions in the previous section, we propose the following rule for a feature extractor on S3DIS:

$$\mathbf{x}_i^{l+1} = \max_{j \in \mathcal{N}(i)} \text{MLP}(\mathbf{x}_j^l \parallel \mathbf{x}_j^l - \mathbf{x}_i^l \parallel \mathbf{p}_j^l - \mathbf{p}_i^l \parallel \mathbf{p}_j^l \parallel \text{dist}(\mathbf{p}_j, \mathbf{p}_i)). \quad (1)$$

In the GNN literature it is well known that incorporating multiple aggregators can boost performance [7]. To assess whether this observation also applies to point cloud data, we also evaluate replacing the max aggregator with combinations of multiple aggregators using the approach proposed by [7]; although the max aggregator provides strong performance when operating on geometric data, it may be sub-optimal when applied to color information. We investigate adding additional min and mean aggregators to assess whether they aid feature extraction.

The results obtained on S3DIS Area 5 are presented in Table 3. Our proposed feature extractor improves performance over the baseline DGCNN model by 2 mIoU, with similar computational overhead. We observe that adding the min aggregator does improve the architecture’s ability to discriminate local geometry, however the increased representational power leads to overfitting rather than generalization; in contrast, we find that the addition of the mean aggregator offers negligible benefit to accuracy overall.

Summary We have demonstrated the value of designing specialized feature extractors when working with point cloud data. Despite using simplified backbones, we have shown that the addition of these feature extractors can noticeably boost accuracy beyond the baseline architecture.

Our conclusions provide new insight for future works seeking to design efficient, but still expressive, architectures operating on point clouds.

5. Further Studies

In this section we present a collection of studies to expand our conclusions from the previous sections. We first show that it is possible to get near state-of-the-art performance on ModelNet40 using Transformers [31] if a good feature representation is provided. We also present studies on how to design simplified feature extraction layers that trade accuracy for reduced resource consumption. Finally, we provide an in-depth analysis of our approach’s robustness and resource consumption.

5.1. Feature Extraction for Transformers

In recent years we have witnessed Transformers achieving state-of-the-art performance across tasks including natural language processing [31], speech recognition [11], and image classification [8]. However, we have yet to see unmodified Transformer models achieving this level of performance on point cloud tasks. Fitting with the hypothesis underpinning this work, we believe that Transformers can only achieve good performance if appropriate features are provided as input to the Transformer layers: raw point positions alone are insufficient. One approach to this issue is to modify the Transformer layers, such that they have only superficial resemblance to the original architecture [40]; however, adopting this approach precludes the usage of advances in Transformer models in the literature—especially with regard to improvements in efficiency. We adopt an alternative approach of using our feature extraction layer, and feeding the resulting features into the Transformer layers.

We use an efficient Transformer variant, the Performer [6], to reduce the memory consumption to linear in the number of points. The model used 4 layers with a dimension of 256 and 4 heads; the feed forward layer dimension was 1024. These parameters were not extensively optimized; we expect with further investigation our results can be improved. The primary aim of this study is to demonstrate the importance of choices in feature extraction, even with little to no fine-tuning.

Our study considers a variety of techniques to feed the points into the Transformer; we consider feeding the raw point positions in after applying a single fully-connected layer (similar to ViT [8]); relative point position encoding (RPPE) from [14], in which each point is encoded with the absolute and relative positions of its nearest neighbors; DGCNN’s feature extractor, and our proposed feature extractor from section 4.1. The results are presented in Table 4. As expected, feeding the raw positional features into the model attains poor performance; RPPE improves upon this by encoding local geometric information into each

points representation, but still achieves relatively poor performance. In contrast, using the DGCNN-based feature extractor (bottom row) yields performance close to that obtainable with a unmodified DGCNN model. As before, incorporating the full set of features in the feature extractor improves performance further. In contrast to our conclusions in section 4.1, it is observed that removing relative point position is most effective for reducing overfitting. One possible explanation for this observation is that the attention mechanism is more effective when each point’s representation also encodes absolute positional information about neighbors; this also explains the effectiveness of RPPE over the raw point positions.

The results in this section provide further support for our hypothesis regarding the importance of feature extraction. One observation we make is that the performance achieved by the Transformer models when combined with the feature extractor layers corresponds closely to the results we achieved in section 4.1. This is further evidence that the feature extraction layer limits the overall model performance.

5.2. Designing Linear-Memory Feature Extractors

The feature extractors we have described in this paper require memory and computation proportional to the number of edges in the constructed graph, as we explicitly construct the message for each edge in the graph as a function of both source and target nodes. While this is acceptable in many cases, it may represent a bottleneck for resource constrained devices; it is also problematic due to the increased data movement which is a significant contributor to power consumption [13]. Another challenge is that message propagation and aggregation must be implemented using index selection operations which are not typically optimized on mobile GPUs or NPUs. As explained earlier, simplified GNN layers can implement these two steps using matrix multiplication-style approaches which are well supported in hardware. These issues motivate us to investigate if we can design feature extractor layers that attain high performance while using $\mathcal{O}(V)$ memory.

This work has already emphasized the importance of the centralization operations in the feature extraction layer. However, it is not possible to use these operations and continue to use matrix multiplication for message propagation around the graph: our procedure to create messages must be a function of the source node only. To this end, we propose the following update rule for a ModelNet40 feature extractor to imitate the centralization operations:

$$\mathbf{x}_i^{l+1} = \left(\max_{j \in \mathcal{N}(i)} \text{MLP}_1(\mathbf{p}_j) \right) - \text{MLP}_2(\mathbf{p}_i). \quad (2)$$

Each point is passed through an MLP, and the resulting representation is subtracted from the result aggregated from

Raw Position	RPPE [14]	Feature Extractor				OA	mAcc
		Source Position	Target Position	Rel Position	Distance		
✓						90.4	86.3
	✓					91.3	87.0
		✓	✓	✓	✓	93.1	88.8
		✓	✓	✓	✓	92.9	89.5
		✓	✓	✓	✓	92.8	88.8
		✓	✓	✓	✓	93.2	89.9
		✓	✓	✓		92.9	89.1
			✓	✓		92.9	89.9

Table 4. Results when combining different feature extractors with a Transformer on ModelNet40; the bottom row corresponds to using a DGCNN-style block for feature extraction. We observe that accuracy improves as the Transformer is combined with more sophisticated feature extractors. Unlike Table 2, we observe that it is better to include source and target positions rather than the relative position. We conjecture that this is due to the query-based attention mechanism used in Transformers.

the local neighborhood. We consider sharing the weights across MLPs, and learning separate MLPs. In principle using separate MLPs should increase model capacity, at the cost of less stable optimization, increased runtime, and potential overfitting.

Experiment	ModelNet40		S3DIS Area 5		
	OA	mAcc	mIoU	OA	mAcc
No centralization	91.7	88.9	49.2	83.3	57.6
Shared MLPs	92.3	89.2	53.5	84.3	61.4
Separate MLPs	93.2	90.6	50.4	84.1	57.4

Table 5. Results when using our resource-efficient feature extraction layer. We observe that adding centralization increases performance on both datasets. Sharing the MLP weights achieves better performance on S3DIS, in contrast to ModelNet40, where adding a separate MLP improves performance beyond the DGCNN baseline. This is due to S3DIS backbone using deeper MLPs in the feature extractor, which is known to induce overfitting and make the optimization less stable [18]

The results when using this feature extractor are shown in Table 5. We provide results using our proposed rule and where no centralization is applied at all. As expected, not using centralization causes noticeable degradation in accuracy. Impressively, our proposed approach surpasses the performance of the full baseline DGCNN model on both ModelNet40 and S3DIS reported in Table 1, although it does not reach the performance of the full feature extraction approach proposed in section 4.1. This is unsurprising given the reduction in model complexity.

Our results indicate that it is better to use separate weights for the MLPs on ModelNet40, and share them on S3DIS. Although this observation is initially counter-intuitive, it is consistent with observations made by Liu et al. [18], where they find that using deeper MLPs in graph layers can impede optimization. The feature extraction layer for the ModelNet backbone uses MLPs of depth 1, in contrast to S3DIS which uses MLPs of depth 2; our ab-

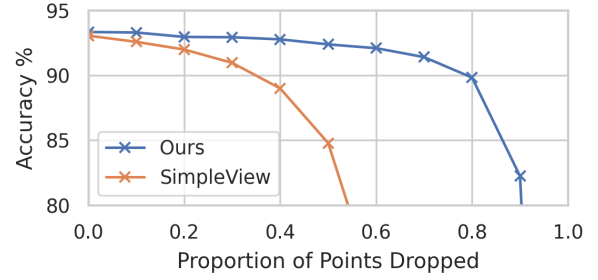


Figure 2. Comparison of our approach with SimpleView, a multi-view CNN, on ModelNet40. Even when 90% of the points have been dropped our approach still obtains accuracy above 80%. Although multi-view methods are easier to optimize, they have significant downsides with regard to robustness.

lated architectures copy the depths faithfully. By sharing the weights, we reduce the potential for overfitting and improve the stability of the optimization, therefore yielding better generalization—even if the theoretical model capacity is reduced.

5.3. Model Robustness and Resource Usage

In the real world, we are also interested in ensuring that our model tolerates imperfect inputs, such as missing points, or high noise. We also care about model latency, memory consumption, and hardware support.

Model Robustness As demonstrated by Goyal et al. [10], multi-view CNNs still represent a strong baseline for point cloud tasks. Given that CNNs have been the target of much research investigating how to make them more efficient, and are well supported in hardware, it could be argued that it makes more sense to optimize this class of models instead of GNNs. In practice, a strong argument against this is that multi-view models are not as robust to variations in point density, or missing points. To illustrate this vi-

Model	ModelNet40		S3DIS	
	CPU	GPU	CPU	GPU
DGCNN	96.3 \pm 2.3	3.00 \pm 0.04	260.0 \pm 1.9	12.31 \pm 0.36
Full Extractor (sec. 4.1)	25.7 \pm 1.1	1.20 \pm 0.05	158.7 \pm 1.8	9.13 \pm 0.06
Full Extractor (sec. 4.1) + SpMM	23.0 \pm 0.9	1.19 \pm 0.06	154.1 \pm 1.9	9.04 \pm 0.13
Efficient Extractor (sec. 5.2) + SpMM	17.7 \pm 0.6	1.09 \pm 0.08	101.3 \pm 1.5	7.71 \pm 0.20

Table 6. Inference times for our models compared to the baseline DGCNN. We observe that our approaches provide especially large speed-ups when using the ModelNet40 backbone, and when running on CPU. Using SpMM provides a small reduction to latency on CPU, in addition to its main benefit of reducing peak memory consumption. Incorporating our efficient feature extractor further reduces model latency.

sually we plot accuracy on ModelNet40 as the proportion of points dropped increases in fig. 2. We include our full model (section 4.1) alongside the SimpleView multi-view CNN model [10]. It can be observed that our models retain high accuracy, even as 95% of points are dropped; in contrast, SimpleView’s accuracy rapidly degrades. We found that deviating from the recommended hyperparameters for SimpleView could enable more robust models to be trained, but with a steep 2% decrease in overall accuracy when evaluating on unmodified point clouds.

Inference Latency In Table 6 we provide inference latency of the baseline DGCNN model, alongside our models which use simplified backbones. We ran our models using PyTorch 1.9; the CPU was an Intel i9-7900x and the GPU was an Nvidia RTX 2080. On S3DIS we are timing the latency to process a single batch of 4096 points. We observe that our approach yields speed-ups of 4.2 \times and 1.7 \times respectively for ModelNet40 and S3DIS on CPU; on GPU we see speed-ups of 2.5 \times and 1.4 \times respectively. Further speed-ups are obtained when the feature extraction layer uses the resource efficient approaches proposed in section 5.2. Although we have achieved significant speed-ups (especially on CPU), we cannot reduce latency further without optimizing other bottlenecks due to Amdahl’s Law; the DGCNN backbone we are modifying has very wide fully connected layers after the graph layers, which contribute a large fraction of the latency. We also note that we have not focused on accelerating the sparse propagation operations in the graph layer: these represent a noticeable bottleneck, especially on GPUs. For a single layer with input and output dimension of 128, we observe speed-ups of 9.9 \times on CPU and 3.5 \times on GPU. To improve end-to-end latency further requires rethinking the model backbone design, which lies beyond the scope of this work.

Memory Consumption For DGCNN models we record peak memory consumption to be 69.7MB on ModelNet40 and 129.6MB on S3DIS. By comparison, our models achieve 19.2MB and 100.3MB peak consumption respec-

tively. Although our simplified layers do achieve a measured 20 \times improvement in memory consumption, once again we do not realise the full benefit of our approach due to the wide fully connected layers that are prevalent in DGCNN’s architecture design.

6. The Road Ahead

Our insight is only one detail with regards to optimizing these models for deployment on resource-constrained devices. Of critical importance is rethinking the backbone architecture design: while our approach has improved the efficiency of these models, we can achieve stronger results by re-designing the backbone with reference to the target hardware implementation.

Going Deeper Most works utilize backbones that are shallow; for example, DGCNN and PointNet++ backbones use 3-4 graph layers. As shown by PosPool [18], going deeper— ≥ 10 graph layers—in combination with a simple layer design can achieve state-of-the-art performance on point cloud tasks. While this strategy is effective for boosting accuracy and retaining low peak memory consumption, it will inevitably increase latency due to the expense of the aggregation operations in GNNs. In contrast, we wish to make the network shallow without sacrificing accuracy.

Increasing Arithmetic Intensity In order to achieve peak accelerator utilization it is vital to achieve high arithmetic intensity: performing many OPs/byte read from memory. Our initial calculations indicate that many elements of the DGCNN architecture achieve arithmetic intensity of just 64 OPs/byte, far below what is required to saturate the a mobile accelerator with limited memory bandwidth. Increasing the memory bandwidth is often not possible, hence we must investigate approaches to increase the OPs/byte. One strategy would be quantization; beyond this, however, requires a novel re-design of our neural network architecture.

6.1. Conclusion

We have investigated PointNet++ and DGCNN, and shown that it is possible to simplify them so long as we retain a sophisticated feature extraction layer. Following our observation, we have shown that by improving the feature extraction layer, we can improve accuracy on ModelNet40 and S3DIS overall despite using resource-efficient backbones. We discuss the improvements to memory consumption and latency our approaches provide; we observe up to 4.5 \times speed-ups on CPUs. We also show that our observation regarding feature extraction also applies to Transformers processing point clouds. Finally, we discuss the next steps required to improve performance of point cloud models on mobile devices.

References

- [1] iPhone 12 Pro. <https://www.apple.com/si/iphone-12-pro/>. Accessed: 28th July 2021.
- [2] Velabit. <https://velodynelidar.com/products/velabit/>. Accessed: 28th July 2021.
- [3] Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding, 2017.
- [4] Yizhak Ben-Shabat, Michael Lindenbaum, and Anath Fischer. 3dmfv: Three-dimensional point cloud classification in real-time using convolutional neural networks. *IEEE Robotics and Automation Letters*, 3(4):3145–3152, 2018.
- [5] Benjamin Charlier, Jean Feydy, Joan Glaunès, François-David Collin, and Ghislain Durif. Kernel operations on the gpu, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74):1–6, 2021.
- [6] Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.
- [7] Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. *arXiv preprint arXiv:2004.05718*, 2020.
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [9] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.
- [10] Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, and Jia Deng. Revisiting point cloud shape classification with a simple and effective baseline. *arXiv preprint arXiv:2106.05304*, 2021.
- [11] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.
- [12] Yu-Xiao Guo and Xin Tong. View-volume network for semantic scene completion from a single depth image. *arXiv preprint arXiv:1806.05361*, 2018.
- [13] M. Horowitz. Computing’s energy problem (and what we can do about it). *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 10–14, 2014.
- [14] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11108–11117, 2020.
- [15] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [16] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31:820–830, 2018.
- [17] Yawei Li, He Chen, Zhaopeng Cui, Radu Timofte, Marc Pollefeys, Gregory Chirikjian, and Luc Van Gool. Towards efficient graph convolutional networks for point cloud handling. *arXiv preprint arXiv:2104.05706*, 2021.
- [18] Ze Liu, Han Hu, Yue Cao, Zheng Zhang, and Xin Tong. A closer look at local aggregation operators in point cloud analysis. *ECCV*, 2020.
- [19] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928. IEEE, 2015.
- [20] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.
- [21] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.
- [22] Haotong Qin, Zhongang Cai, Mingyuan Zhang, Yifu Ding, Haiyu Zhao, Shuai Yi, Xianglong Liu, and Hao Su. Bipointnet: Binary neural network for point clouds. *arXiv preprint arXiv:2010.05501*, 2020.
- [23] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3577–3586, 2017.
- [24] Naai-Jung Shih, Pei-Huang Diao, Yi-Ting Qiu, and Tzu-Yu Chen. Situated ar simulations of a lantern festival using a smartphone and lidar-based 3d models. *Applied Sciences*, 11(1), 2021.
- [25] Siddharth Srivastava and Gaurav Sharma. Exploiting local geometry for feature and graph construction for better 3d point cloud processing with graph neural networks, 2021.
- [26] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015.
- [27] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12):2295–2329, 2017.
- [28] Shyam A Tailor, Javier Fernandez-Marques, and Nicholas D Lane. Degree-quant: Quantization-aware training for graph neural networks. *arXiv preprint arXiv:2008.05000*, 2020.
- [29] Shyam A Tailor, Felix L Opolka, Pietro Liò, and Nicholas D Lane. Adaptive filters and aggregator fusion for efficient graph convolutions. *arXiv preprint arXiv:2104.01481*, 2021.
- [30] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J

- Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6411–6420, 2019.
- [31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [32] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- [33] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.
- [34] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2019.
- [35] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, Jan 2021.
- [36] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes, 2015.
- [37] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.
- [38] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3173–3182, 2021.
- [39] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102, 2018.
- [40] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. *arXiv preprint arXiv:2012.09164*, 2020.
- [41] Yiren Zhao, Duo Wang, Xitong Gao, Robert Mullins, Pietro Lio, and Mateja Jamnik. Probabilistic dual network architecture search on graphs. *arXiv preprint arXiv:2003.09676*, 2020.