Unsupervised Contrastive Representation Learning for 3D Mesh Segmentation (Student Abstract)

Ayaan Haque[1,2],Hankyu Moon[2], Heng Hao[2], Sima Didari[2], Jae Oh Woo[2], Patrick Bangert[2]

[1]University of California, Berkeley, Berkeley, CA, USA
[2]Samsung SDS Research America, CA, USA
ayaanzhaque@berkeley.edu

Abstract

3D deep learning is a growing field of interest due to the vast amount of information stored in 3D formats. Triangular meshes are an efficient representation for irregular, non-uniform 3D objects. However, meshes are often challenging to annotate due to their high computational complexity. Therefore, it is desirable to train segmentation networks with limited-labeled data. Self-supervised learning (SSL), a form of unsupervised representation learning, is a growing alternative to fully-supervised learning which can decrease the burden of supervision for training. Specifically, contrastive learning (CL), a form of SSL, has recently been explored to solve limited-labeled data tasks. We propose SSL-MeshCNN, a CL method for pre-training CNNs for mesh segmentation. We take inspiration from prior CL frameworks to design a novel CL algorithm specialized for meshes. Our preliminary experiments show promising results in reducing the heavy labeled data requirement needed for mesh segmentation by at least 33%.

Introduction

3D polygonal meshes are efficient at representing objects with irregular and non-uniform surfaces. The pioneering method for working directly on meshes is MeshCNN (Hanocka et al. 2019) which designs mesh-specific convolutional and pooling layers. Traditionally, training deep neural networks requires large datasets, but labeling meshes is challenging for various reasons. Thus, training with partially labeled datasets is desirable. Unsupervised representation learning (URL), self-supervised learning (SSL), and contrastive learning (CL) are a family of training frameworks which have been successfully applied to addressing the limited-labeled data, or small data, problem. URL enables a network to learn strong visual representations without any supervision. In SSL, a form of URL, a supervisory signal is produced synthetically from the unlabeled data itself. In CL, a network learns to sort latent representations based on similarity. Since CL is a form of SSL, we label our method as both SSL and CL. There have been many recent successful works on CL (He et al. 2020; Caron et al. 2020; Chen et al. 2020). SimCLR(Chen et al. 2020) is a standout CL method which many works build upon. There have been limited works which apply URL to meshes (Zhou, Bhatnagar, and Pons-Moll 2020; Zimmermann, Argus, and Brox 2021). However, URL methods for segmentation as well as SimCLR-based CL methods for mesh analysis remain under explored. We introduce self-supervised MeshCNN, or SSL-MeshCNN, a mesh-specialized CL method to perform downstream segmentation with limited-labeled data.

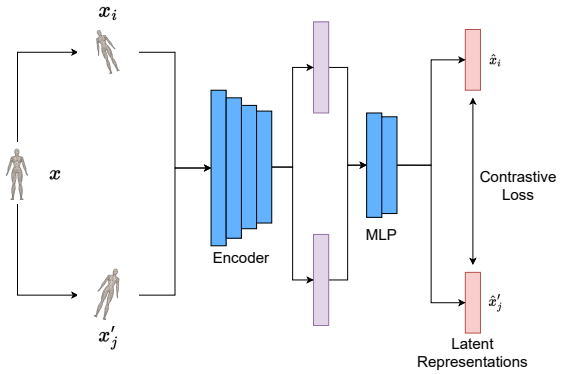

Figure 1: Schematic of our contrastive learning framework.

Methods

A triangular mesh can accurately represent surfaces and topology of objects. A mesh is described by its vertices, edges, and faces. Using conventional image-based CNNs for meshes is infeasible as they are irregular and non-uniform. MeshCNN (Hanocka et al. 2019) trains directly on meshes by utilizing mesh-specific convolutional and pooling layers. Each edge in a mesh is used to create a 5-dimensional input feature set. Our overall training procedure involves two steps: CL pre-training using an entire dataset without labels (shown in Figure 1), followed by downstream segmentation using only samples with corresponding labels.

CL learns efficient representations by maximizing agreement between two augmented versions of the same input in the latent space. For each input x in minibatch of size M_1, using augmentation function Aug $(\cdot)$, we generate two uniquely augmented meshes $x_i = \text{Aug}(x)$ and $x'_j = \text{Aug}(x)'$, forming positive pair (x_i, x'_j). Each batch now has $2M_1$ samples and $2(M_1 - 1)$ negative samples. For CL, our architecture is a MeshCNN encoder followed by a nonlinear two-layer

Dataset Proportion (%)	Segmentation Accuracy (%)		
	w/o CL	w/ CL	Diff
5	58.02	63.15	5.15
10	63.89	66.29	2.40
25	81.36	83.61	2.25
33	83.98	85.08	1.10
50	85.25	87.84	2.59
67	86.69	88.97	2.28
75	87.52	90.35	2.83
100	88.58	90.50	1.92

Table 1: Segmentation accuracy with and without CL pre-training at varying quantities of labeled data.

MLP head. Contrastive loss is used to maximize agreement between representations of positive pairs and minimize agreement between representations of negative pairs. After pre-training, the encoder is saved and the MLP is discarded.

To perform SimCLR-based CL, a strong augmentation policy is required to create effective positive and negative pairs. SimCLR uses image augmentations, but mesh augmentations differ significantly, so the augmentation policy must be redesigned. This is primarily because our input features are similarity-invariant, meaning augmentations such as isotropic scaling, rotation, and translation will not change the feature input (Hanocka et al. 2019). We randomly apply a series of three augmentations: anisotropic scaling, vertex shifting, and edge flipping. Hanocka et al. (2019) reports these specific augmentations are effective for meshes. Anisotropic scaling scales each vertex of a face to random degrees. Vertex shifting rearranges random vertices to different locations. Edge flipping flips edges between two adjacent faces. The augmentation strengths are stochastically tuned during training.

We use NT-Xent (Chen et al. 2020) as our contrastive loss. Once CL is complete, we transfer the pre-trained encoder to a Mesh-UNet, which is trained on a limited quantity of meshes x and labels y to predict semantically segmented meshes $\hat{y}$ using standard cross-entropy loss.

Experimental Evaluation

For experimentation, we use the Human Body Segmentation dataset (Maron et al. 2017), an 8-class segmentation task (381 training, 18 testing meshes). For CL, we use the entire training set without labels. For downstream segmentation, we use varying quantities of samples and labels.

Table 1 displays the segmentation accuracy at varying portions of labeled data with and without CL. Our results demonstrate that CL improves performance over no CL baselines at all dataset portions. At the lowest levels of supervision (5%), we report a five percentage point increase in accuracy due to CL. With CL, the network matches fully-supervised performance when trained on just 67% of examples. Thus, our CL framework reduces the supervision requirement by 33%. Due to random sampling in experiments, in use cases, any 67% of samples can be labeled to achieve our result.

Figure 2 displays visualized segmentation results of our network at varying quantities of labeled data with and without contrastive pre-training. As shown, when pre-trained with CL, the network achieves more accurate and representative segmentations than the fully-supervised baselines. The borders between classes are more distinct and do not bleed into each other when training with CL, further confirming the superiority of our contrastive pre-training.

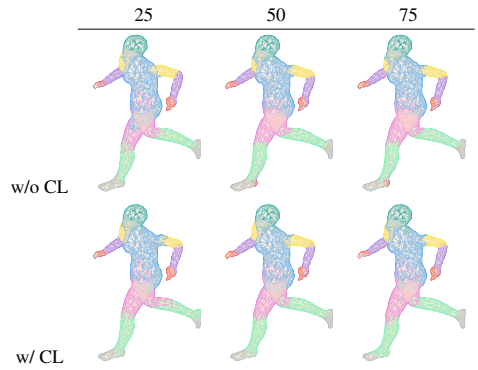

Figure 2: Predicted segmentation visualizations at varying proportions of dataset (%) with and without CL pre-training.

Conclusions

We have presented SSL-MeshCNN, a tailored CL augmentation policy for 3D meshes, providing the network with efficient learned representations to perform downstream segmentation with reduced supervision. Our preliminary results confirm the effectiveness of our method at learning strong representations, which reduces the need for labeled examples for mesh segmentation by at least 33%. Our future work will focus on designing more rigorous augmentation policies.

References

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33: 9912–9924.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Hanocka, R.; Hertz, A.; Fish, N.; Giryes, R.; Fleishman, S.; and Cohen-Or, D. 2019. Meshcnn: a network with an edge. *ACM Trans. Graph.*, 38(4): 1–12.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Maron, H.; Galun, M.; Aigerman, N.; Trope, M.; Dym, N.; Yumer, E.; Kim, V. G.; and Lipman, Y. 2017. Convolutional neural networks on surfaces via seamless toric covers. *ACM Trans. Graph.*, 36(4): 71–1.

Zhou, K.; Bhatnagar, B. L.; and Pons-Moll, G. 2020. Unsupervised shape and pose disentanglement for 3d meshes. In *European Conference on Computer Vision*, 341–357. Springer.

Zimmermann, C.; Argus, M.; and Brox, T. 2021. Contrastive representation learning for hand shape estimation. In *DAGM German Conference on Pattern Recognition*, 250–264. Springer.