Exploring Multiview Pooling for Deep 3D Shape Descriptors

Jaeseong Lee Brady Zhou

Introduction

The emergence of large scale 3D model datasets has motivated the need for a more compact representation of 3D objects for applications such as classification and retrieval. In this work, we choose to represent objects with a feature vector based off of various 2D projections from different viewpoints, similar to Su et. al [1], and experiment with different methods of pooling views using information entropy.

The motivation behind this work was given a collection views, does there exist a subset of these views that provide substantial information to describe the image?

Related Work

The main inspiration behind this work is MVCNN by Su et. al [1], in which they render an object from 80 views, feed the 2D images into a fine-tuned image convolutional neural network, and max-pool the feature vectors in the later fully connected layers to create a multi-view descriptor. This approach performs surprisingly effectively on the popular ModelNet40 dataset.

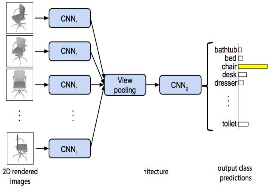

Figure 1. Multiview CNN

Wu et. al [2] chose a different representation for 3D objects, using a voxel based approach with 3D CNNs and an architecture based off generative adversarial networks.

Brock et. al [3] also chose a voxel based representation, and using Inception and residual layers achieve state of the art classification on ModelNet40 with 95.54%.

Technical Details

In this work, we use the ModelNet10 dataset [2] to evaluate performance on the classification task. The dataset consists of 10 object categories, with a total of 5000 objects, represented as triangle meshes. We pick 25 viewpoints to render the object from, and these are used as inputs into our model.

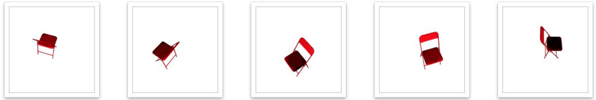

Figure 2. Sample views used in our model

Each step of training consists of the following steps.

We render the object from several views, and feed each image into a CNN.

We try to follow the architecture of [1] as closely as possible before the pooling layers to generate a similar feature space.

Each of these feature vectors are then pooled in various methods.

The Baseline Model: MVCNN

Each mesh is associated with a set of rendered image $\{x_1, x_2, \cdots, x_n\}$, which are fed into a CNN to generate a set of feature vecto $\{f_1, f_2, \cdots, f_n\}$ These are then combined to form a single descriptc $z^i = max(f_1^i, f_2^i, \cdots, f_n^i)$.

Shannon Informational Entropy

$$\mathrm{H}(X) = \sum_{i=1}^{n} \mathrm{P}(x_i)\,\mathrm{I}(x_i) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i),$$

We sort the feature vectors by their information score, and pick the top K best views to perform max pooling on.

Feature Similarity Distance Score

$$distance(f_i, f_{global}) = ||f_i - f_{global}||$$

We perform the max pooling operation on the feature vectors and pick the top K nearest neighbors (L2 norm) and re-pool using only those vectors.

Combined Score

We define the combined score of a image to be a linear combination of the informational entropy term and the feature similarity.

Image Global Saliency Model

This model is trained on a regression task, where the labels are a handcrafted function of the relative scores for a single mesh.

After this model was trained, the pooled feature vector is a linear combination of these normalized scores with the original feature vectors.

$$z = \sum_{i=1}^{n} w_i f_i \qquad w_i = \frac{e^{s_i}}{\sum_{k=1}^{n} e^{s_k}}$$

Results

Figure 3. Sample predicted "good" views. Left is good, right is bad.

Method	Accuracy
MVCNN Baseline	82.3
Top-1 Entropy	79.8
Top-5 Entropy	86.1
Top-10 Entropy	**89.8**
Combined Entropy-Distance	82.0

Table 1. Accuracy table with entropy score model.

To improve our models, we could incorporate the learning of weights for the combined score instead of setting them empericaly. In this work we defined saliency to be informational entropy, which could be too much of an assumption. The informational entropy is defined by the model's confidence in the predictions, which is a learned combination of feature vectors and these features could be very different than an observer's notion of saliency.

References

1. Su, Hang, et al. "Multi-view convolutional neural networks for 3d shape recognition." *Proceedings of the IEEE international conference on computer vision*. 2015.
2. Wu, Jiajun, et al. "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling." *Advances in Neural Information Processing Systems*. 2016.
3. Brock, Andrew, et al. "Generative and Discriminative Voxel Modeling with Convolutional Neural Networks." *arXiv preprint arXiv:1608.04236* (2016).