Contrastive-GMMs for Open-World Learning

Stephen Scarano, Preston Yee, Kobe Falus

Abstract

We present a light-weight, deep contrastive network which utilizes a Gaussian mixture model to identify class instances unseen at train time. Our model compares favorably with the state-of-the-art on ModelNet40, and we perform an ablation study to determine the impact of our newly-introduced Gaussian loss term on both the learned embeddings and model performance. Last, we discuss the potential left un-analyzed in our empirical work here, and provide direction for future work.

1. Introduction

The most successful and popular deep learning projects have traditionally been those tested in a strictly closed-setting; that is, the model is evaluated with respect to some number of classes observed during training [8]. This paradigm—which is also temporally-frozen—is the subject of much scrutiny in recent literature, such that its tendency towards domain shift [11], opaqueness [10], and overfitting [1] are well-documented. Our approach rejects the closed-set assumption, identifying novel instances not seen at train time. Further, similar work out of the zero-shot learning community has seen massive performance gains leveraging Large Language Models (LLMs) in their embeddings space; however, this comes at the cost of model size [9, 14]. Our model is comparably much smaller while also not requiring any retraining of the deep network as novel classes are introduced. This is ideal, since deep models trained continuously suffer the well-known trouble of catastrophic forgetting – that is, prior information is rapidly lost [8]. Indeed, we seek to construct a model that is at least partially resistant to these frictions while sporting competitive performance.

2. Related Work

Our approach benefits from a cross-disciplinary corpus at the intersection of contrastive learning, 3D-shape classification, zero-shot learning, and open-world learning literature. Our project goal is most aligned with the latter two communities, whose the stated goal of identifying classes unseen at training time motivates our work, though we clarify their different focuses: a zero-shot approach can be considered a subset of the open-world paradigm, and largely neglects the role of time in learning [4]. While work specific to the 3D setting is sparse, recently the zero-shot community has leveraged Large-Language models (LLMs) to great effect on classification and segmentation tasks [14] [9].

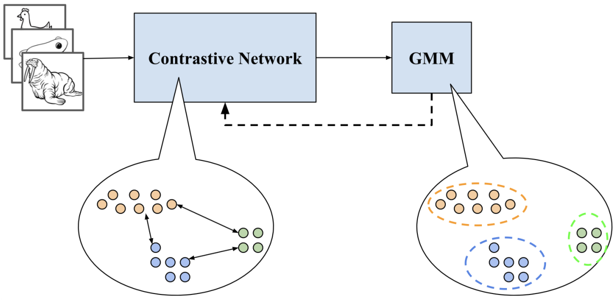


Figure 1. Our model architecture decomposed into ”Contrastive” and ”GMM” blocks.

The Open-World Learning literature is nearly vacant with respect to 3D data; however, we adapt many of its conventions for analyzing known and unknown data classes. For instance, we implement an accuracy metric described by [4] to scale priority given to novel instances vs. known instances; that is, our reported accuracy (NA) is a mixture of the novel (N) and known (K) accuracy metrics:

$$NA = \lambda_r(ACC_K) + (1 - \lambda_r)(ACC_N)$$

Additionally, we adopt a revised contrastive-learning pipeline popularized by [2] which integrates augmented instances into training. The authors use this perturbed data to minimize the distance between feature embeddings of the same class. [6] adapts this pipeline for labeled data, which are usually found to bolster performance [3].

3. Methodology

We illustrate our pipeline in Figure 1, which we decompose into a ”Contrastive Block” and ”Gaussian Mixture Model (GMM) Block”. At a high level, our contrastive block learns embeddings for each image which maximize distances between foreign class instances and minimize distances between same-class instances. The results are fed into the GMM block, which fits the embeddings and provides the log-liklihood as feedback. The loss term, then, is simply a sum of the superivsed contrastive loss (see [6]) and the GMM log-liklihood:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{CTR} + \mathcal{L}_{GMM} \\ &= \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp\left(z_i \cdot z_p/\tau\right)}{\sum_{a\in A(i)} \exp\left(z_i \cdot z_a/\tau\right)} + \mathcal{L}_{GMM}\end{aligned}$$

Where $\mathcal{L}_{CTR}$ is the contrastive loss, $\mathcal{L}_{GMM}$ is the Gaussian loss, and τ is a temperature hyperparameter. In practice, however, we scale the Gaussian-term with a constant δ to identify an ideal trade-off between the Gaussian and contrastive objectives. That is,

$$\mathcal{L} = \mathcal{L}_{CTR} + \delta\mathcal{L}_G MM$$

The contrastive block is comprised of four convolutional blocks; that is, four groupings of **(a)** convolutional layer, **(b)** layer-norm, **(c)** leaky ReLU (negative slope 0.01), **(d)** max-pool, and **(e)** a second convolutional layer. Our pipeline primarily augments image data using *RandomCrop* and *RandomJitter* perturbations as recommended by the authors of [2].

We adopt the Multi-View CNN (MVCNN) approach from [12] to adapt our process to 3D data. Our deep network is trained under the contrastive and Gaussian objectives described above *per each 2D view* of each instance, where each instance is provided 80 unique views. At test-time (or validation), we compute the mean of these 80 embeddings and feed-forward the resulting global descriptor into a Gaussian Mixture Model (GMM).

3.1. Prediction

Before deployment, we fit a GMM to a sample of the known class embeddings, which avoids perturbation from novel instances. Incoming data is then processed by the deep network and input to the GMM for prediction. We apply the mahalanobis distance as a natural point threshold, as its acknowledgement of distribution covariance better suits Gaussians than Euclidean distance. We express the mahalanobis distance below:

$$d(x_1, x_2) = \sqrt{(x_1 - x_2)^T C^{-1} (x_2 - x_2)}$$

Here, C is the covariance matrix and $x_1, x_2 \in \mathbb{R}^n$ are points in the distribution. An ideal threshold (expressed as a percentile of all sampled distances) is determined empirically. Those exceeding the chosen threshold are labeled as novel instances.

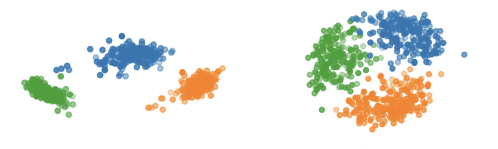
4. Experiments

Our experimentation largely focuses on novelty detection, since there exist comparable benchmarks and baselines at our disposal. Our setting for both sections 4.1 and 4.2 are conducted on a subset of *ModelNet40* [13]—chosen for its unique and contrasting classes—under a setting of 3 known classes and 1 unknown class. We note that this does not take full advantage of the Open-World paradigm, since we do not discriminate between novel instances, which remains out of scope for our work here. We measure model success by the lambda-accuracy discussed in section 2, as it provides the reader an intuition for the trade-off between our known-class and novel-class accuracy metrics. Typically, we will describe the accuracy for all $\lambda \in \{0, 0.25, 0.50, 0.75, 1\}$.

4.1. Ablation Studies

δ	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$
0	1	0.85	0.71	0.58	0.43
0.25	**0.7**	**0.72**	**0.74**	**0.76**	**0.78**
0.5	0.7	0.7	0.71	0.71	0.72

Table 1. Lambda Accuracy by GMM Loss constant (δ)



(a) without GMM Loss (b) with GMM Loss

Figure 2. 3D-embeddings after 10 epochs of training with and without the GMM loss term.

We examine the impact of the GMM loss term both qualitatively and quantitatively: Figure 2 show our model embeddings with and without the GMM loss term. Embeddings trained without the Gaussian loss term are, expectedly, less gaussian, which we confirm empirically in Table 1. Both findings suggest a trade-off between the Gaussian and contrastive objectives, where the proper δ is selected empirically over the validation set, optimizing for $\lambda = 0.5$. Here we select $\delta = 0.25$ according to this criteria, and use this model for the remainder of the paper.

4.2. Baseline Comparison

We compare our model results to that of **(a)** confidence thresholding, a simple but general open-set classification method, and **(b)** Clip2Point, a recent adaptation of the zero-shot method CLIP to 3D [5, 7]. We implement thresholding from scratch, training another MV-CNN with an identical architecture but capped with a softmax layer. We then threshold the output probabilities $p^{(i)}$ by some t such that if $p^{(i)} < t$ we label the instance as novel.

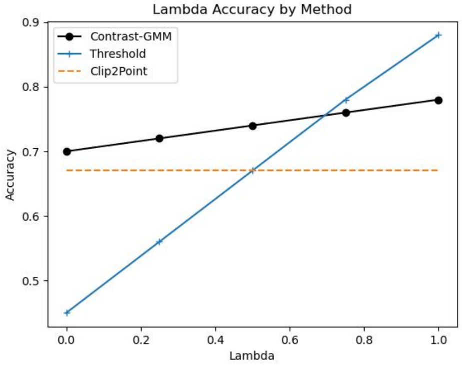


Figure 3. Lambda-accuracy (NA) by method

Figure 3 shows the results of our experiment, which are compelling, since we can achieve comparable accuracy to Clip2Point; however, we re-emphasize that these results cover a limited space of the Open-World problem. We observe instances of only one novel class, and thus do not share the burden of reconciling rival novel classes as Clip2Point does, since this was out of the scope for this paper.

5. Limitations

A few things limited the scope and quality of our study:

1. Baseline models for novelty detection are not common. Not many projects look at novelty classification, and even more so for an open-world setting. We had to augment pre-existing projects to focus on novelty classification, and this may have biased our results.

2. The lack of resources, specifically GPUs, became a problem when trying to emulate other studies that required it to run. Some of us had computers that did not have GPUs, so we couldn't run them on our own computers. When modifying some of these models to not run on GPUs, they were very slow. Our model was difficult to run, so we had to use Google Colab T40, which wasn't ideal due to memory limits.

3. Due to our time constraints, we were not able to do as much testing as we wanted. With more time we would have been able to use a larger dataset for testing and would have been able to prepare more/better baselines. We also were not able to explore as much as we would have like open world data, since it was not feasable to run our model on those scenes in time.

6. Conclusions

This paper pairs contrastive learning and Gaussian mixture models to classify data unseen at train time with a lightweight deep architecture. Our ablation trials appear to motivate the introduction of our Gaussian loss term, and the model appears to perform well in comparison to state-of-the-art zero-shot classifiers. Though we are pleased with the current results, we stress that the greatest advantages of the model are left untested, which motivates future work:

1. We do not examine the model's capability to discriminate between novel classes — this is problematic, since a major question of Open-World methods is whether the semantic information learned from known classes is sufficiently discriminative for unknown classes.

2. Related to (1) we perform no lifelong learning analysis of the model; that is, we do not investigate how long we can delay retraining the contrastive network by leveraging the Gaussian mixture model.

3. Related to both (1) and (2), this project was conducted under limited time and compute conditions, and though our method would likely benefit from a greater number of known classes and data instances, here it is evaluated in a limited setting.

At a high-level, we believe that for deep learning to integrate harmoniously into everyday life, it must be prepared for its realities: a closed setting fixed in time is not adequate to solve problems in a dynamic world characterized by unpredictable phenomena.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. 1

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. 1, 2

[3] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners, 2020. 1

[4] Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3614–3631, oct 2021. 1

[5] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson W. H. Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training, 2022. 3

[6] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. 1, 2

[7] Thomas C.W. Landgrebe, David M.J. Tax, Pavel Paclík, and Robert P.W. Duin. The interaction between classification and reject performance for distance-based reject-option classifiers. *Pattern Recognition Letters*, 27(8):908–917, 2006. ROC Analysis in Pattern Recognition. 3

[8] Martin Mundt, Yong Won Hong, Iuliia Pliushch, and Visvanathan Ramesh. A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning. *CoRR*, abs/2009.01797, 2020. 1

[9] Songyou Peng, Kyle Genova, Chiyu ”Max” Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies, 2023. 1

[10] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019. 1

[11] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. A closer look at domain shift for deep learning in histopathology, 2019. 1

[12] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proc. ICCV*, 2015. 2

[13] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes, 2015. 2

[14] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip, 2021. 1