Self-Supervised Pretraining with Momentum Contrast (MoCo) for Out-of-Distribution Detection

Subhasish Bandyopadhyay
January, 2025

***Abstract*—Out-of-distribution (OOD) detection is a critical aspect of deploying robust and reliable machine learning models. This paper investigates the efficacy of self-supervised pretraining using Momentum Contrast (MoCo) for OOD detection. We present a comprehensive pipeline where a ResNet-18 model is pretrained on the CIFAR-10 dataset using MoCo, followed by fine-tuning for classification. The pretrained and fine-tuned model is then evaluated for its OOD detection capabilities on the Street View House Numbers (SVHN) dataset. We employ three distinct OOD detection metrics: Mahalanobis distance, energy scores, and a novel hybrid Energy-Gradient Score (EGS). Our findings demonstrate that energy-based methods, particularly the EGS, significantly outperform the Mahalanobis distance in distinguishing between in-distribution and out-of-distribution data. This suggests that self-supervised pretraining with MoCo can learn rich and generalizable features that are highly effective for OOD detection, and that the choice of OOD detection metric plays a crucial role in leveraging the full potential of these learned representations.**

***Index Terms*—Out-of-Distribution Detection, Self-Supervised Learning, Momentum Contrast, Energy-Gradient Score, Mahalanobis Distance, Energy Score**

I. Introduction

The remarkable success of deep learning models in a wide range of applications has been accompanied by a growing concern about their reliability and safety in real-world scenarios. One of the key challenges in this regard is the problem of out-of-distribution (OOD) detection. OOD detection refers to the task of identifying whether a given input is from the same distribution as the training data or from a different, unseen distribution. This is of paramount importance in safety-critical applications such as autonomous driving, medical diagnosis, and financial fraud detection, where a models inability to recognize and handle OOD inputs can have severe consequences. Traditionally, deep learning models have been trained under the closed-world assumption, where the test data is assumed to be drawn from the same distribution as the training data. However, in real-world deployments, models are often confronted with a plethora of OOD inputs, ranging from novel object categories to adversarial attacks. A standard deep learning model, when presented with an OOD input, is likely to produce a high-confidence but incorrect prediction, which can be misleading and dangerous. Therefore, it is crucial to equip models with the ability to not only make accurate predictions on in-distribution data but also to express uncertainty and identify OOD inputs. In recent years, self-supervised learning has emerged as a powerful paradigm for learning rich and generalizable representations from large amounts of unlabeled data. Self-supervised learning methods, such as Momentum Contrast (MoCo), learn by creating pretext tasks where the model is trained to predict certain properties of the data itself, without the need for human-annotated labels. This allows the model to learn the underlying structure and semantics of the data, resulting in features that are more robust and transferable to a wide range of downstream tasks, including OOD detection. This paper explores the use of MoCo for OOD detection. We hypothesize that the features learned through MoCo pretraining, which are not biased towards a specific set of classes, will be more effective in distinguishing between in-distribution and out-of-distribution data compared to features learned through traditional supervised training. To test this hypothesis, we implement a pipeline where a ResNet-18 model is pretrained on CIFAR-10 using MoCo, and then fine-tuned for classification. We then evaluate the OOD detection performance of the fine-tuned model on the SVHN dataset. To assess the OOD detection capabilities of our model, we employ three different metrics:

1) Mahalanobis Distance: A classical statistical distance measure that has been successfully applied to OOD detection in deep learning models.
2) Energy Score: A more recent and highly effective method that uses the energy of the models output logits to distinguish between in-distribution and out-of-distribution samples.
3) Energy-Gradient Score (EGS): A novel hybrid score that combines the energy score with the norm of the input gradient, aiming to capture both the models confidence and the sensitivity of its features to input perturbations.

Through a series of experiments and comprehensive analysis, we aim to provide insights into the effectiveness of MoCo pretraining for OOD detection and to compare the performance of different OOD detection metrics in this context.

II. Methodology

Our experimental methodology is designed to systematically evaluate the effectiveness of MoCo pretraining for OOD detection. The pipeline consists of several stages, including data loading and augmentation, model architecture definition, MoCo pretraining, fine-tuning and classifier training, and OOD detection.

A. Datasets and Data Loading

We use two datasets in our experiments:

- CIFAR-10: This dataset serves as our in-distribution dataset. It consists of 60,000 32x32 color images in 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images.
- Street View House Numbers (SVHN): This dataset is used as our out-of-distribution dataset. It consists of 32x32 color images of house numbers, with a total of 99,289 images.

We use different data augmentation techniques for the pretraining and fine-tuning stages. For MoCo pretraining, we apply a series of strong augmentations to generate different views of the same image, which is essential for contrastive learning. These augmentations include random resized cropping, random horizontal flipping, color jittering, and random grayscale conversion. For the fine-tuning and classifier training stages, we use a similar set of augmentations, but with slightly different parameters. For the test and OOD datasets, we only apply normalization to the images.

B. Model Architecture

Our model architecture consists of two main components: a MoCo encoder and a classifier head.

- MoCo Encoder: We use a ResNet-18 architecture as the base encoder for our MoCo model. The MoCo model itself consists of two encoders: a query encoder and a key encoder. Both encoders have the same ResNet-18 architecture. The query encoder is updated through gradient descent, while the key encoder is updated as an exponential moving average of the query encoder. This momentum-based update mechanism is a key feature of MoCo and helps in maintaining a consistent and large dictionary of negative keys for contrastive learning. The output of the encoders is a 128-dimensional feature vector.
- Classifier Head: On top of the MoCo encoder, we add a simple multi-layer perceptron (MLP) classifier head. The classifier head takes the 128-dimensional feature vector from the encoder as input and consists of a linear layer, a batch normalization layer, a ReLU activation function, a dropout layer, and a final linear layer that outputs the logits for the 10 classes of CIFAR-10.

C. MoCo Pretraining

The MoCo pretraining is the core of our self-supervised learning pipeline. The goal of this stage is to learn rich and generalizable visual representations from the unlabeled CIFAR-10 training data. The pretraining is done using the InfoNCE contrastive loss function, which is defined as:

$$L_{\text{MoCo}} = -\log \frac{\exp(q \cdot k_+/\tau)}{\exp(q \cdot k_+/\tau) + \sum_{j=1}^{K} \exp(q \cdot k_j^-/\tau)} \quad (1)$$

where q is the query embedding, k_+ is the positive key embedding, k_j^- are the negative key embeddings, and τ is a temperature parameter. The pretraining is performed for 300 epochs with a batch size of 256. We use the Adam optimizer with a learning rate of 0.001 and a weight decay of 10^{-5}.

D. Fine-Tuning and Classifier Training

After the MoCo pretraining, we fine-tune the encoder and train the classifier head on the labeled CIFAR-10 training data. This stage consists of two phases:

1) Fine-Tuning: In the first phase, we train both the encoder and the classifier jointly for 10 epochs.
2) Classifier Training: In the second phase, we freeze the encoder and train only the classifier head for another 20 epochs.

For both phases, we use the standard cross-entropy loss function and the Adam optimizer. We also employ a cosine annealing learning rate scheduler and early stopping based on the validation loss.

E. Out-of-Distribution Detection Methods

Once the model is trained, we evaluate its OOD detection performance using three different metrics.

1) Mahalanobis Distance: The Mahalanobis distance measures the distance of a feature vector f from the mean of the in-distribution features μ under a covariance Σ:

$$D_M(f) = \sqrt{(f-\mu)^\top \Sigma^{-1} (f-\mu)} \quad (2)$$

A higher Mahalanobis distance indicates that the sample is more likely to be OOD.
2) Energy Score: The energy score is based on the models output logits z:

$$E(z) = -T \cdot \log \sum_{i=1}^{C} \exp\left(\frac{z_i}{T}\right) \quad (3)$$

where T is a temperature parameter. A higher energy score indicates a higher likelihood of the sample being OOD.
3) Energy-Gradient Score (EGS): The EGS is a hybrid score that combines the normalized energy score $\hat{E}$ with the norm of the input gradient:

$$\text{EGS} = \alpha \cdot \hat{E} + (1-\alpha) \cdot \|\nabla_x \|f\|_2\|_2 \quad (4)$$

where α is a weighting factor.

III. Results

The results of our experiments demonstrate the effectiveness of MoCo pretraining for OOD detection and provide a clear comparison of the performance of the different OOD detection metrics.

A. OOD Detection Performance

We evaluated the OOD detection performance using the Area Under the Receiver Operating Characteristic (AUROC) curve and the False Positive Rate at 95% True Positive Rate (FPR@95TPR). The results are summarized in Table I.

TABLE I: OOD Detection Performance Metrics

Metric	AUROC	FPR@95TPR
Mahalanobis Distance	0.2579	0.9973
Energy Score	0.9106	0.3446
Energy-Gradient Score (EGS)	0.9506	0.2281

B. Visualization of OOD Scores

To further visualize the separation between in-distribution and out-of-distribution samples, we plotted histograms of the OOD scores for each metric. Figure 1 shows a significant overlap between the distributions of the in-distribution (CIFAR-10) and out-of-distribution (SVHN) samples for the Mahalanobis distance, indicating that this metric is not very effective in separating the two. In contrast, the energy-based scores provide much clearer separation. The histogram for the Energy-Gradient Score (EGS), shown in Figure 2, demonstrates the best separation among the three metrics. The distribution of the OOD samples is concentrated in the high-score region, resulting in a distinct separation from the in-distribution samples. The standard Energy Score (histogram not shown) also provides good separation, though not as pronounced as the EGS.

C. ROC Curve Analysis

We also plotted the ROC curves for the three OOD detection metrics to compare their performance. The ROC curves confirm the results from the histograms and the performance metrics. The EGS curve is consistently above the other two curves, indicating its superior performance across all thresholds. The energy score curve is also significantly better than the Mahalanobis distance curve, which performs close to random chance.

IV. Discussion

The results of our experiments provide strong evidence for the effectiveness of self-supervised pretraining with MoCo for OOD detection. The high AUROC and low FPR@95TPR values achieved by the energy-based methods, particularly the EGS, demonstrate that the features learned through MoCo are highly discriminative and can be effectively used to distinguish between in-distribution and out-of-distribution data. The poor performance of the Mahalanobis distance in this context is noteworthy. This can be attributed to the nature of the feature space learned by contrastive methods like MoCo. These methods aim to spread features across the surface of a hypersphere to maximize contrast, rather than clustering them into compact, class-conditional Gaussian distributions. This violates the core assumption of the Mahalanobis distance. This dispersion and overlap are visualized in the t-SNE embedding of the features shown in Figure 4, where the in-distribution (CIFAR-10) and out-of-distribution (SVHN) features are heavily intermingled. On the other hand, energy-based methods, which are based on the models output logits, are less sensitive to the geometric properties of the feature space and are more directly related to the models confidence in its predictions. This makes them more suitable for OOD detection with models pretrained using MoCo. The superior performance of the EGS over the standard energy score suggests that incorporating information about the sensitivity of the models features to input perturbations can further improve OOD detection performance. The gradient norm term in the EGS acts as a regularizer that penalizes models with highly sensitive features, which are more likely to be associated with OOD inputs.

V. Conclusion

In this paper, we have demonstrated the effectiveness of self-supervised pretraining with Momentum Contrast (MoCo) for out-of-distribution (OOD) detection. Our experiments on the CIFAR-10 and SVHN datasets show that the features learned through MoCo are highly effective for distinguishing between in-distribution and out-of-distribution data. We have also shown that energy-based methods, particularly the novel Energy-Gradient Score (EGS), are more suitable for OOD detection with MoCo-pretrained models compared to the traditional Mahalanobis distance. Our findings have several important implications for the field of OOD detection and self-supervised learning. First, they highlight the potential of self-supervised learning as a powerful tool for building more robust and reliable machine learning models. Second, they underscore the importance of choosing the right OOD detection metric to fully leverage the benefits of self-supervised pretraining. Future work could explore the use of other self-supervised learning methods for OOD detection, as well as the development of new and more effective OOD detection metrics that are specifically designed for self-supervised learned representations.

References

[1] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[2] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems*, 2018, pp. 7167–7177.

[3] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Advances in Neural Information Processing Systems*, 2020, pp. 21464–21475.

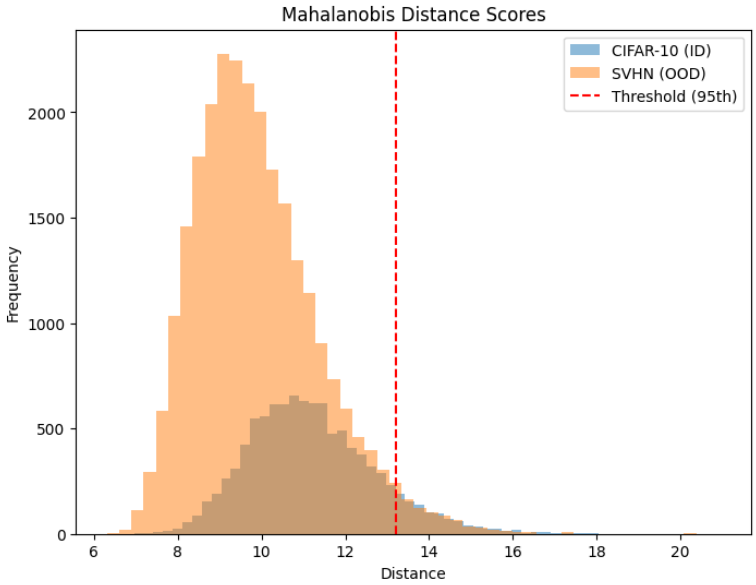

Fig. 1: Histogram of Mahalanobis distance scores for indistribution (CIFAR-10) and out-of-distribution (SVHN) data. The significant overlap indicates poor separation.

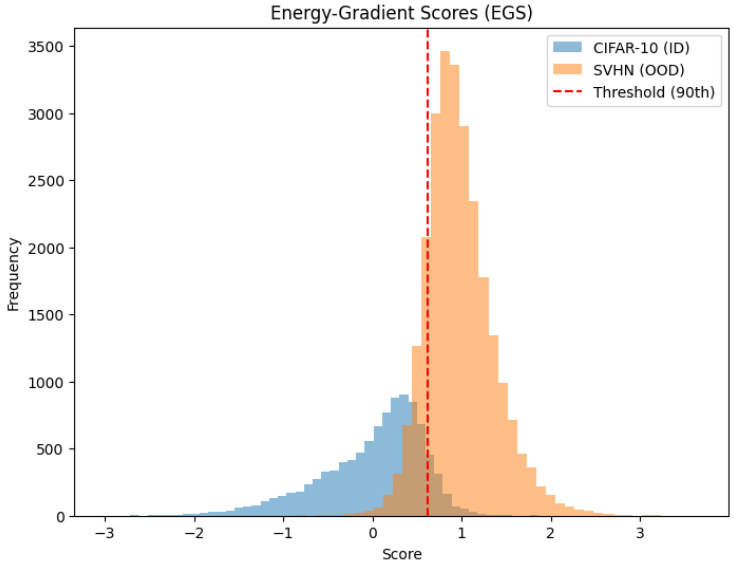

Fig. 2: Histogram of Energy-Gradient Scores (EGS). The separation between ID and OOD data is much clearer compared to the Mahalanobis distance.

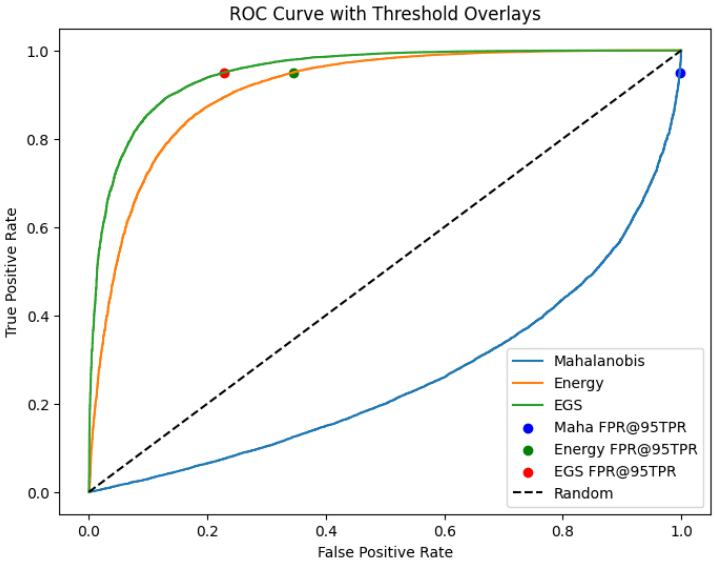

Fig. 3: ROC curves for Mahalanobis, Energy, and EGS methods. The EGS curve is closest to the top-left corner, indicating superior performance.

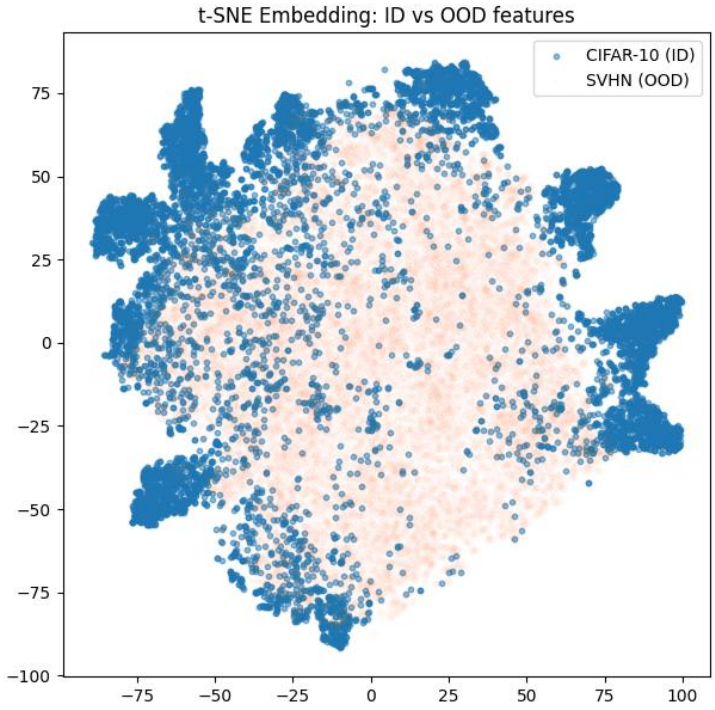

Fig. 4: t-SNE visualization of feature embeddings for ID (CIFAR-10) and OOD (SVHN) data. The lack of clear separation in the feature space helps explain the failure of the Mahalanobis distance.