

Thermodynamics-inspired OOD Detection

Salmaan Barday
University of Cape Town

Gary Louw
University of Cape Town
With AI Safety South Africa & Apart Research

Abstract

We propose a block level extension to the energy based out-of-distribution (OOD) detector of Liu et al. [1], computing and aggregating energy scores at each ResNet18 residual block, while also performing a temperature sweep to find optimal temperature for different datasets.

In experiments on mixed OOD datasets, softmax confidence edges out free energy when In and Out-of-distributions are very similar, but our free energy method greatly outperforms the softmax confidence the more they differ. We also discover that the optimal temperature rises with OOD distance. These results show that the block level free energy scoring could potentially be useful and motivates further cross-dataset benchmarking.

Keywords: Out-of-distribution detection, thermodynamics.

Introduction

Out-of-distribution (OOD) detection is necessary for the safe deployment of AI systems in public settings. The authors in [1] investigate a technique for OOD detection inspired by thermodynamics. They consider a discriminative neural classifier ϕ mapping an input $\mathbf{x}$ to an array of output logits $\mathbf{z} = (z_1(\mathbf{x}), \dots, z_C(\mathbf{x}))$. For each class c , [1] defines an *energy*

$$E_c(\mathbf{x}) = -T \log \frac{\exp(z_c(\mathbf{x})/T)}{\sum_{k=1}^C \exp(z_k(\mathbf{x})/T)}, \quad (1)$$

which is marginalised over classes to obtain the *free energy*

$$F(\mathbf{x}; T) = -T \log \sum_{c=1}^C \exp(z_c(\mathbf{x})/T), \quad (2)$$

where T is the *temperature*. Our extension of [1] is to hook into each block of the ResNet to calculate a block-wise energy $E_\ell(\mathbf{x})$. Boltzmann weights are assigned and pooled across all blocks:

$$F_{\text{total}}(\mathbf{x}) = \sum_{\ell=1}^L w_\ell E_\ell(\mathbf{x}). \quad (3)$$

By varying T , one adjusts the influence of high-energy layers in the free energy score. The result is a multi-depth signal for “novelty” in OOD detection.

Motivations

OOD detection allows a neural net to flag input it cannot reliably parse. Reliable OOD methods are therefore essential for any application where the output of the model has important real-world implications (e.g., medical diagnoses). Sources in the literature (e.g., [1]) suggest that thermodynamics-inspired approaches might outperform traditional methods using the softmax function. We investigate and extend the technique of [1] in this project. Specifically, we extract energy scores from each ResNet block to quantify how much an input deviates from the patterns seen during training, where larger block-level energies indicate greater “outlierness” at that scale. We then aggregate these multi-scale energy measurements and sweep the temperature parameter to establish thresholds that reliably flag truly novel (i.e., out-of-distribution) samples.

Setup and Results

In-Distribution Benchmark. We used the CIFAR-10 dataset as our in-distribution benchmark. CIFAR-10 comprises 60000 color images of resolution 32×32 pixels across 10 object classes. We followed the standard split of 50000 images for training and 10000 for testing, and reserved 5000 of the training images (10%) as a validation set. All images were normalized to zero mean and unit variance per channel.

Out-of-Distribution Test Sets. To assess OOD detection performance, we used two external datasets. First, *SVHN* (Street View House Numbers), a collection of 99289 RGB digit images cropped from real world, street view photos, representing a far OOD domain. Second, *CIFAR-100*, which contains 60000 images of the same size and acquisition conditions as CIFAR-10 but across 100 object classes, representing a near OOD domain. For each dataset, we concatenated the train and test splits to form two independent OOD test sets.

Model and Training. We extend the framework of Liu et al. [1] by computing energy scores at each of the L residual blocks in a ResNet-18 backbone. In contrast, the original method derives free energy solely from the final pre-softmax logits.

We fine-tuned ResNet-18 on CIFAR-10 for 10 epochs using SGD with momentum and a variance-based regularizer on the logits to encourage smoother activations. Forward hooks

captured pre-softmax activations from each block, from which we computed block-wise energy scores $E_\ell(\mathbf{x})$. These were aggregated with learned Boltzmann weights w_ℓ into a total free energy score. We swept the temperature parameter T over a logarithmic grid in $[0.1, 40000]$ on the CIFAR-10 validation set to select the T maximizing ROC-AUC.

Evaluation Protocol. We evaluated OOD detection on five splits in addition to the in-distribution test set. First, each of the two pure OOD sets: SVHN and CIFAR-100. Then, five mixed OOD splits created by sampling equal numbers of SVHN and CIFAR-100 images and combining them at the following ratios:

12.5% CIFAR-100 + 87.5% SVHN
 25% CIFAR-100 + 75% SVHN
 50% CIFAR-100 + 50% SVHN
 75% CIFAR-100 + 25% SVHN
 87.5% CIFAR-100 + 12.5% SVHN

For each split, we computed free energy using the selected temperature T , as well as the softmax baseline. We then measured area under the receiver operating characteristic curve (ROC-AUC) for each metric by labeling CIFAR-10 test images as “in” and OOD images as “out.”

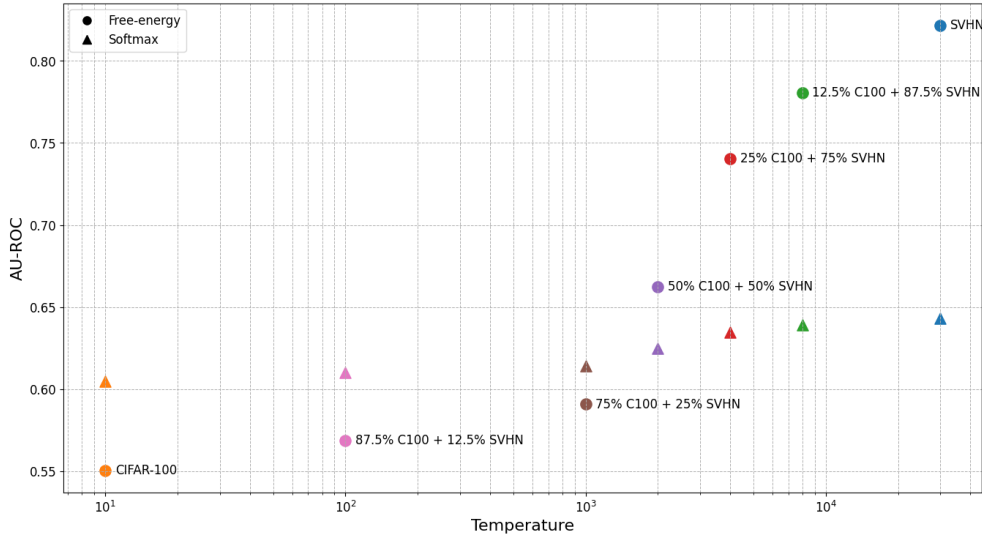


Figure 1: Area under the ROC curve (AU-ROC) for out-of-distribution detection plotted against the temperature parameter (log scale). Circles denote the thermodynamics inspired free energy score, while triangles denote the softmax baseline. Note that the same colour indicates shapes are from the same dataset. Results are shown for two pure OOD domains, CIFAR-100 and SVHN as well as for five mixed splits combining CIFAR-100 and SVHN. For each split, the optimal temperature was selected by maximizing the free energy ROC-AUC on the validation set, then applied at test time. This plot illustrates how detection performance varies with dataset similarity and temperature tuning.

Discussion and Conclusion

We observe a performance crossover between softmax confidence and free energy scoring as the OOD data shifts from resembling the training set to diverging from it. When the OOD split is dominated by CIFAR-100 (small domain gap), softmax confidence slightly outperforms free energy. As more SVHN data is added (larger domain gap), free energy takes the lead and widens its margin.

The optimal temperature also increases with OOD distance. Although both methods improve at higher temperatures, free energy gains much more than softmax. This suggests tuning to higher temperatures and relying on free energy scoring for datasets with strong novelty.

Because free energy’s advantage depends on how similar the OOD data are to the training set, it would be useful to repeat the temperature sweep on other dataset pairs. Some that are closely aligned with the training data, others very different, to confirm whether this crossover trend holds more generally.

Finally, benchmarking our block level extension against the original Liu et al. [1] method on the same splits and comparing both energy based approaches to softmax confidence and other state of the art OOD detectors, would provide a comprehensive view of their relative strengths and weaknesses.

References

- [1] W. Liu, Y. Pang, D. Xu, *et al.*, “Energy-based Out-of-distribution Detection,” in *NeurIPS Proceedings*, 2021.
<https://arxiv.org/pdf/2010.03759>