

1 Abstract

Stochastic-sampling-based Generative Neural Networks, such as Restricted Boltzmann Machines and Generative Adversarial Networks, are now used for applications such as denoising, image occlusion removal, pattern completion, and motion synthesis. In scenarios which involve performing such inference tasks with these models, it is critical to determine metrics that allow for model selection and/or maintenance of requisite generative performance under pre-specified implementation constraints. In this paper, we propose a new measure for quantifying generative model performance based on $p$-values derived from the combined use of Maximum Mean Discrepancy (MMD) and permutation-based (PT-based) resampling, which we refer to as the PT-MMD metric. We demonstrate the effectiveness of this metric for two cases: (a) Selection of bitwidth and activation function complexity to achieve minimum power-at-performance for Restricted Boltzmann Machines; (b) Quantitative comparison of images generated by two types of Generative Adversarial Networks (PGAN and WGAN) to facilitate model selection in order to maximize the fidelity of generated images. For both these applications, our results are shown using both Euclidean and Haar-based kernels for the PT-MMD test. This demonstrates the critical role of distance functions in comparing generated images versus their corresponding ground truth counterparts as what would be perceived by human users.

2 Introduction and Proposed Methodology

Generative Models refer to models of probability distributions that can be used to codify high dimensional data. Evaluating and comparing the performance of such models is an area of significant interest [2]. MMD has been proposed as an evaluation and training metric for GANs [5,9]. We propose a novel methodology where we construct hypothesis tests by combining MMD and PT-based resampling to be used as an evaluation measure for inference tasks on pre-trained generative models. The $p$-values obtained from our PT-MMD based tests serve as a metric/measure that can be utilized for tasks such as comparing quality within a family of models, e.g. GANs, and facilitating selection of hyperparameters, such as bitwidths and sigmoidal activation function complexity, for performing inference on hardware realizations of generative neural networks, e.g. RBMs.

A novel aspect of our proposed framework is addressing the limitation of conventional MMD kernels by comparing the performance of PT-MMD with both Euclidean and Haar-transform based kernels for evaluating image quality. It has been shown that Euclidean distance based metrics used in many generative machine learning frameworks have perceptual limitations [15]. The Haar-transform based distance metric has been shown to be strongly correlated with human opinion scores of image quality when compared to Euclidean distance [7][11]. We construct our MMD kernel by using the directional components of the Haar-transform and demonstrate the advantage of this kernel versus Euclidean based kernels in our PT-MMD hypothesis test framework for both RBMs and GANs.

---

1 Bitwidth is specifically known to be linearly correlated with power consumption in RBMs [13].
Our hypothesis testing framework is built on permutation-based resampling of data. This approach is advantageous as it does not require a priori knowledge of the null distribution, which for finite samples is dataset specific, hence the desire to estimate it with an empirical approximation [6]. In addition, the resampling approach is well-suited for performing inference in real-time environments where the available amount of data may be limited. The following sections detail our PT-MMD methodology and results for inference tasks using RBMs and GANs.

3 RBM-MNIST

3.1 Methods

RBMs are a class of probabilistic generative models that have been used in applications such as image classification, sequence completion, and speech recognition [8]. When implemented on low-power digital processors, applying suitable metrics to network design for obtaining power-at-performance is critical. We demonstrate the use of PT-MMD in selecting bitwidth and sigmoidal function complexity in a RBM trained on the MNIST dataset [4]. As shown in Fig. 1 we consider three sigmoid approximations of varying complexity suitable for hardware implementations [12].

Figure 1: This figure demonstrates the hardware sigmoids being considered in this experiment. We compare them to the true sigmoid. Note that as fidelity increases, so does hardware implementation cost [12, 13].

Given a user-specified significance level \( \alpha \), our design goal is to determine the optimal sigmoid and bitwidth, i.e. to achieve optimal power consumption, while at the same time meeting performance constraints which can be specified as significance levels for the test.

3.2 Results

Figure 2a shows the results of our PT-MMD test using the Euclidean-distance kernel for different bitwidths and sigmoid realizations. It can be seen that the PLAN sigmoid gives the highest \( p \)-values across for bitwidths of 16 and higher which is consistent with the fact that PLAN achieves the lowest error uniformly among all sigmoid approximations considered for our analysis as shown in Figure 1b. In the case of the Haar distance kernel, PLAN still gives the best results however in this case near-zero \( p \)-values are seen for both Ramp and AS. Based on these results it can be concluded that the PLAN approximation with a 16-bit implementation is the lowest bitwidth implementation which can generate MNIST images of sufficient quality with high similarity to the original distribution. In addition, given the fact that Haar distance is a perceptual metric, the PT-MMD test based on this kernel is able to better distinguish synthetic images generated using lower quality sigmoids like AS from the ground truth as seen in the contrast between \( p \)-values produced by AS with Euclidean versus Haar-based PT-MMD.

4 GAN-LSUN

4.1 Methods

GANs are an important class of Deep Generative Models that when trained on large image datasets have been shown to reproduce images of extremely high quality. Various types of GANs have been proposed in the literature using

\[ \text{Such significance levels can be based on pre-determined human opinion scores of image quality for given datasets.} \]
**Figure 2:** p-value results for RBM Generated MNIST Data are shown above with the user specified significance level of 0.05 as a demonstrative example. In this application, the higher the p-value, the better the quality of the synthetic distribution. The trials are run at bitwidths of 4, 8, 12, 16, 24, 32, 48, and 64. A total of 100 Monte Carlo simulations were used for each bitwidth to capture the variation in its p-values.

Different architectures. For such networks, quantifying differences in generative performance based on rational metrics for critical considerations such as model selection is an area of significant interest. We consider two GANs, PGAN and WGAN, trained on the LSUN dataset and apply our PT-MMD based tests to the generated images from these models [1, 13, 10, 14]. Sample data generated from these GANs is shown in Figure 3 subjectively demonstrating their generative quality differences. Quantification of the performance of these two models is inefficient to perform manually. Therefore, the goal is to automate this process using PT-MMD.

**4.2 Results**

We apply our PT-MMD test to compare GAN generated LSUN-like images against the LSUN ground truth database using both the Euclidean and Haar-based MMD kernels.

<table>
<thead>
<tr>
<th></th>
<th>Euclidean</th>
<th>Haar</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGAN</td>
<td>0.2316 ± 0.0811</td>
<td>0.00527 ± 0.00539</td>
</tr>
<tr>
<td>WGAN</td>
<td>0.00 ± 0.00</td>
<td>0.00 ± 0.00</td>
</tr>
</tbody>
</table>

These results are given in Table 1. The WGAN fails the PT-MMD test at any significance level for both types of kernels. In contrast, the PGAN achieves higher p-values using the Euclidean distance kernel. In case of Haar, the observed p-values are non-zero which allows the user to differentiate the PGAN model versus the WGAN.
Figure 4: Cumulative Distribution Function (CDF) of obtained MMD scores in PT-MMD runs. For each graph, a decrease of the degree of separation in the two estimated CDFs corresponds to an increase in mean \( p \)-value of the PT-MMD test which signifies increasing closeness between ground truth and synthetic data distributions. Notably, under Haar PT-MMD, the separation between the two CDFs is much higher than under Euclidean PT-MMD, suggesting that the Haar distance is better at distinguishing synthetic distributions from true distributions.

The differences in PT-MMD performance are illustrated in more detail in Figure 4 which shows the separation between the baseline MMD values of the synthetic data and the MMD values obtained via permutations to estimate the null distribution. It can be seen that the extent of separation of the baseline and permutation CDFs is much higher for the PGAN resulting in higher \( p \)-values for the PT-MMD test as compared to the WGAN.

5 Conclusions

In this paper, we quantify the performance of generative models consistent with image quality assessment as performed by human users. This is done by constructing a novel permutation variant of the 2-sample kernel MMD test based on Euclidean and Haar distances. The applicability of PT-MMD is demonstrated for enabling determination of bitwidth and sigmoid approximation, which are critical design parameters for RBM hardware implementations. Further, we show that PT-MMD facilitates automated model selection between two GANs. For the datasets considered we have also shown that the perceptual Haar distance based PT-MMD is better at distinguishing images, both RGB and grayscale, versus PT-MMD based on Euclidean distance.
References