LEARNABLE K-SAMPLING TO ENHANCE DEEP LEARNING MODEL ROBUSTNESS AND GENERALIZABILITY

Dary Lu
Department of Electrical and Computer Engineering
Duke University
Durham, NC 27708
dl370@duke

ABSTRACT

Deep learning (DL) methods have shown promise in medical image processing. However, due to the challenges posed by privacy regulations and the scarcity of labeled clinical data, researchers have to rely on publicly available neuroimaging datasets to train the model. Thus performance in real-world clinical settings is often limited by the lack of robustness and generalizability. In this study, we investigate the impact of differing image quality and scanner types on the performance of DL models for magnetic resonance imaging (MRI) data. To address these issues, we propose the use of K-sampling to reduce the features fed into the model, enhancing both its robustness and generalizability. We develop and optimize a K-sampling layer and evaluate its effectiveness in improving the performance of deep learning models on diverse datasets with varying image quality. Our findings demonstrate the potential of K-sampling as a valuable tool to address the challenges of robustness and generalizability in deep learning models for medical image processing, paving the way for more effective deployment in real-world clinical scenarios.

1 Introduction

Deep learning (DL) methods have demonstrated remarkable potential in the field of medical image processing, outperforming traditional algorithms in various tasks [1]. As data-driven algorithms, deep learning models excel when provided with substantial amounts of data, enabling them to achieve superior performance. Consequently, numerous models are being considered for practical applications in real-world clinical settings. In these contexts, deep learning models are expected to efficiently handle diverse types of data.

Nonetheless, acquiring extensive datasets for training DL models, particularly in magnetic resonance imaging (MRI), poses significant challenges due to privacy regulations and the scarcity of labeled clinical data. Consequently, researchers frequently rely on publicly available neuroimaging datasets [2]. These datasets, however, have inherent limitations, such as images acquired from a single scanner with consistently high image quality, which does not accurately represent real-world clinical scenarios. First, the high performance of deep learning models on one dataset cannot be readily transferred to unseen external datasets [3]. Second, ghosting artifacts also influence model performance [4]. As a result, models trained on data from one type of scanner with exceptional image quality often struggle to perform well on more diverse datasets.

One potential solution to enhance model robustness and generalizability involves processing the input images to reduce the features fed into the model. In this context, robustness refers to the model's ability to perform well on varying image qualities, while generalizability concerns its performance on previously unseen datasets. [5] concludes that features related to the fibrograndular tissue of the breast are more sensitive to scanner parameters than the features related to tumor only. In this case, by reducing the input features, the model is more likely to learn the essential characteristics of the common label. In this study, we employ K-sampling to diminish the MR data features.

In our project, we investigate the impact of differing image quality and scanner types on model performance. Furthermore, we develop and optimize a K-sampling layer to address the challenges of robustness and generalizability in deep learning models for medical image processing.

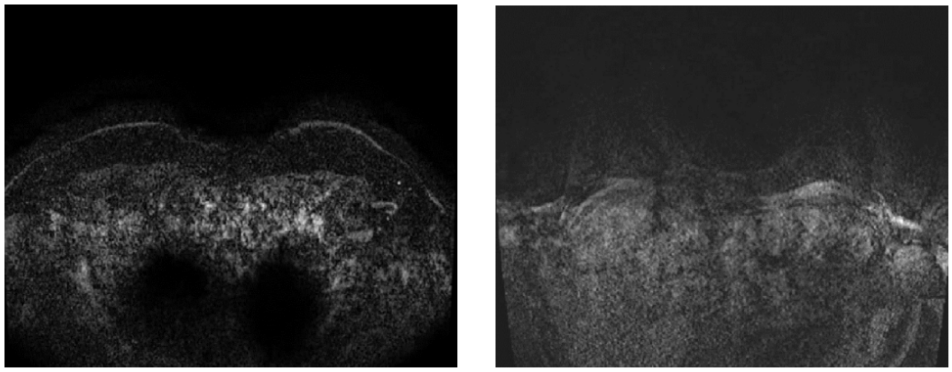

Figure 1: GE breast MRI image vs. Siemens breast MRI image.

2 Related Work

A substantial body of research has been dedicated to the application of deep learning methods in medical image processing. Many of these studies [1, 4] have primarily focused on conducting comprehensive surveys of deep learning techniques and their performance in various scanners and image qualities. Moreover, some existing work has attempted to address the challenges in this domain. A significant portion of these efforts has been aimed at improving the model structure itself to enhance the performance of deep learning models, often involving modifications to the architecture, loss functions, or optimization techniques to increase accuracy. For example, Kyathanahally et al. [6] used a fully connected network to deal with MR spectra containing spurious echo ghost signals, while Yan et al. [7] employed domain adaptation to handle data across different scanners.

Building upon previous work, our study focuses on incorporating a physically-inspired K-sampling layer into the deep learning model. By integrating this layer, we aim to enhance its robustness and generalizability across varying image quality and scanner types. This approach presents a novel contribution to the field, as it specifically targets the challenges that arise from the limited availability and diversity of training data, which has been a persistent issue in applying deep learning methods to medical image processing.

3 Methods

3.1 Data

The dataset used in this project is the Duke-Breast-Cancer-MRI [8], which consists of 773,888 breast MRI scans acquired using four types of 1.5T and 3T clinical-grade MR scanners. These scans were collected from 922 patients treated at Duke Hospital between January 1, 2000, and March 23, 2014, who had invasive breast cancer and available pre-operative MRI examinations. The primary goal of this collaboration is to test MRI's potential in prognosticating patients' short and long-term outcomes, as well as predicting pathological and genomic features of the tumors.

In this project, we specifically used data from GE (General Electric) and Siemens scanners. The sample image can be viewed in Fig 1. Our model is designed for classification purposes, and the image labels indicate only whether a tumor is present or not, without considering the tumor type or stage.

3.2 K-Sampling Layer and Network

Our proposed model follows a pipeline that includes the following stages: (1) conversion of the input image to its K-space representation using Fast Fourier Transform (FFT), (2) application of the K-sampling method, (3) conversion of the K-sampled data back to the image domain using Inverse Fast Fourier Transform (IFFT), and (4) processing the resulting image with the AlexNet convolutional neural network (CNN) architecture. In this section, we will introduce the implementation of the K-sampling method and briefly discuss the AlexNet architecture. The model diagram can be viewed in Fig 2.

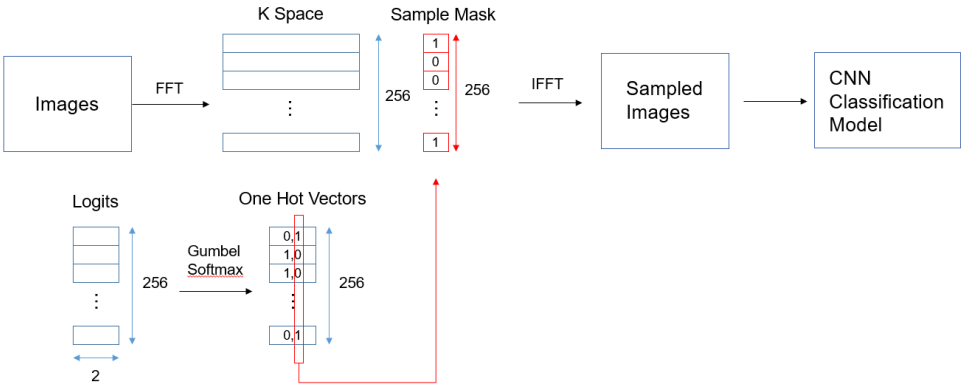

Figure 2: Model Network: It consists a K-Sampling layer and a CNN classification model

Due to the incompatibility of standard thresholding or direct sampling methods with backpropagation, we employed the Gumbel Softmax technique. We generated logits for each K-space vector and used Gumbel Softmax to create one-hot vectors. In detail, we initialized the logit as:

$$Logit(0) = \log \frac{Rate}{1 - Rate} \tag{1}$$

where the logit is a $1X2$ vector with $Logit(1) = 0$, and the Rate is the sampling rate. This logit is used to simulate the sample probability.

Then, we generated one-hot vectors using the hard Gumbel Softmax.[9] Gumbel Softmax is an algorithm that uses softmax to simulate the sampling process:

$$Z = onehot\left(\max i \mid \pi_1 + \ldots + \pi_{i-1} \leq U\right) \tag{2}$$

where π represents the logits.

It first uses this formula to sample. This is a "reparameterization trick", refactoring the sampling of Z into a deterministic function of the parameters and some independent noise with a fixed distribution.

$$Z = onehot\left(argmax_i\ G_i + \log\left(\pi_i\right)\right) \tag{3}$$

where $G_{\mathrm{i}} \sim Gumbel(0, 1)$ are i.i.d. samples drawn from the standard Gumbel distribution.

Then, softmax is used as a differentiable approximation to argmax. As $\tau \to 0$, the softmax computation smoothly approaches the argmax, and the sample vectors approach one-hot; as $\tau \to \infty$, the sample vectors become uniform. In this case, the whole process has a gradient now. The sample vectors y are now given by

$$y_i = \exp\left(\left(G_i + \log\left(\pi_i\right)\right)/\tau\right) / \Sigma_j \exp\left(\left(G_j + \log\left(\pi_j\right)\right)/\tau\right)$$

for every $i = 1, \ldots, x.\tau$ is the temperature parameter that controls how closely the new samples approximate discrete, one-hot vectors. Next, we selected the second column to form the sample mask, allowing the autodiff (automatic differentiation) to optimize the logits.

Due to the computation limits, we use AlexNet as the CNN model in this project is AlexNet[7], which is a popular architecture. It consists of multiple convolutional and pooling layers, followed by fully connected layers and a softmax output layer. In our model, we utilized the AlexNet architecture to process the IFFT output from the K-sampling method, allowing the model to learn and classify the presence of tumors in the MRI scans.

3.3 experiment Process

In our experiments, we evaluated the impact of ten different K-sampling rates on our model's performance. Our experimental framework consisted of data split, train, and test phases. As mentioned in the data section, we had two

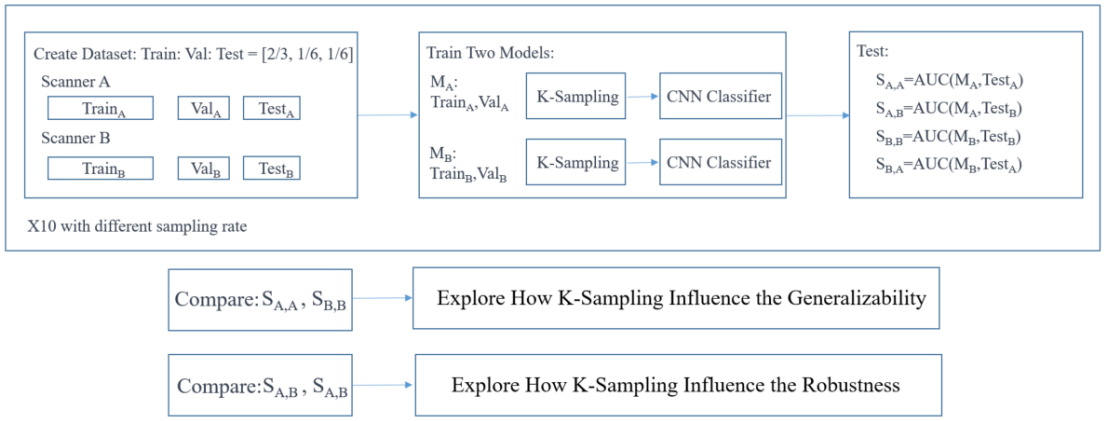

Figure 3: experiment Process: Build two models based on two different dataset, and test them on different test data

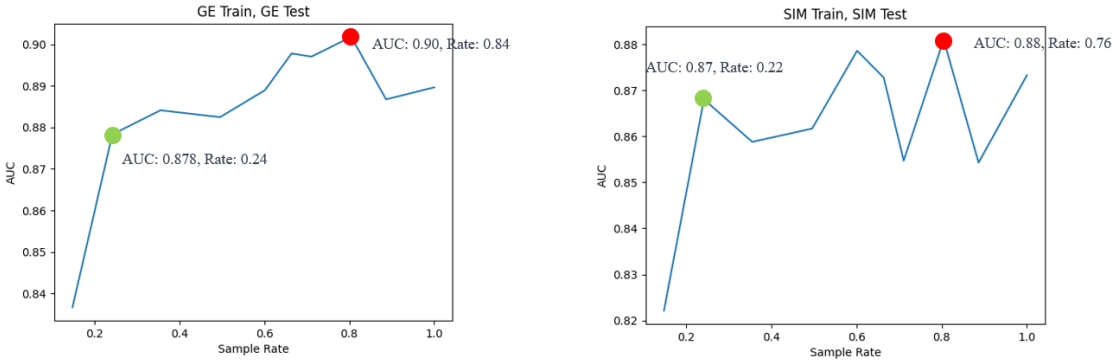

Figure 4: Robust Compare Curve: Green points are the best robustness and the red points are the best AUC

datasets, A and B, which corresponded to data from GE (General Electric) and Siemens scanners. We then trained two models, m_A and m_B, based on these datasets A and B, respectively. After that, we get the $S_{i,j}$ by use test data i to test Model j.

To assess the robustness and generalizability of our models, we conducted the following tests:

(1): Robustness was evaluated by viewing the performance scores $S_{A,A}$ and $S_{B,B}$. A higher score with a low sampling rate indicated good robustness in handling variations within the same scanner type. (2): Generalizability was assessed by viewing the performance scores $S_{A,B}$ and $S_{B,A}$. A higher score indicated good generalizability in handling unseen data from different scanner types.

By analyzing the results of these tests, we aimed to demonstrate the how K-sampling layer influences the robustness and generalizability of deep learning models. Then, based on the result, we will propose the optimized layer.

4 Results

Regarding robustness, it is surprising that when the sampling rate is around 0.24 for GE-trained models and 0.22 for Siemens-trained models, both achieve their maximum robustness. In fact, the AUC values for GE and Siemens models at these sampling rates are 0.878 and 0.868, respectively. This indicates that we can select a limited number of K-space vectors and still maintain performance close to the optimal one. In this case, if the sampling rate is low, the model is more stable when dealing with different-quality images, as it only needs a few vectors in K-space. As for generalizability, models with lower sampling rates exhibit higher AUC, potentially because they learn common features across the scanners. For example, when the sampling rate is 0.148, the AUC for the GE-trained model tested on Siemens data is 0.689, while the AUC for the Siemens-trained model tested on GE data is 0.726. We can see that

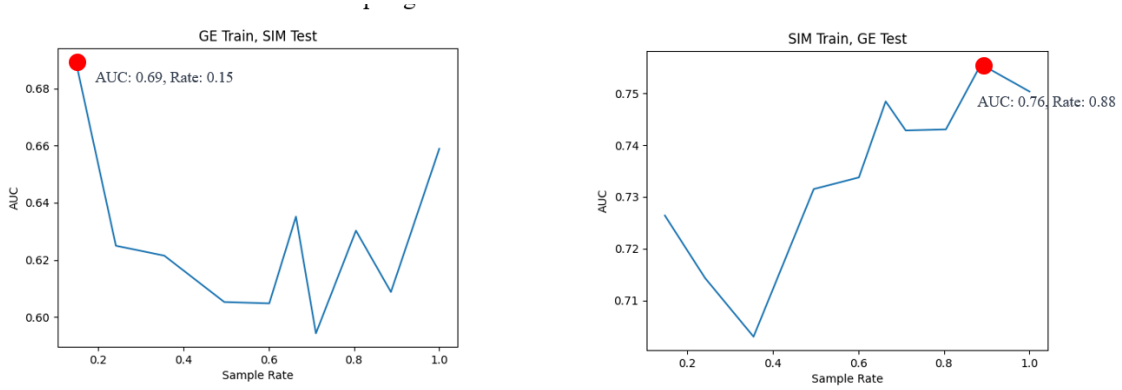

Figure 5: Generalizability Compare Curve: the red points are the best AUC

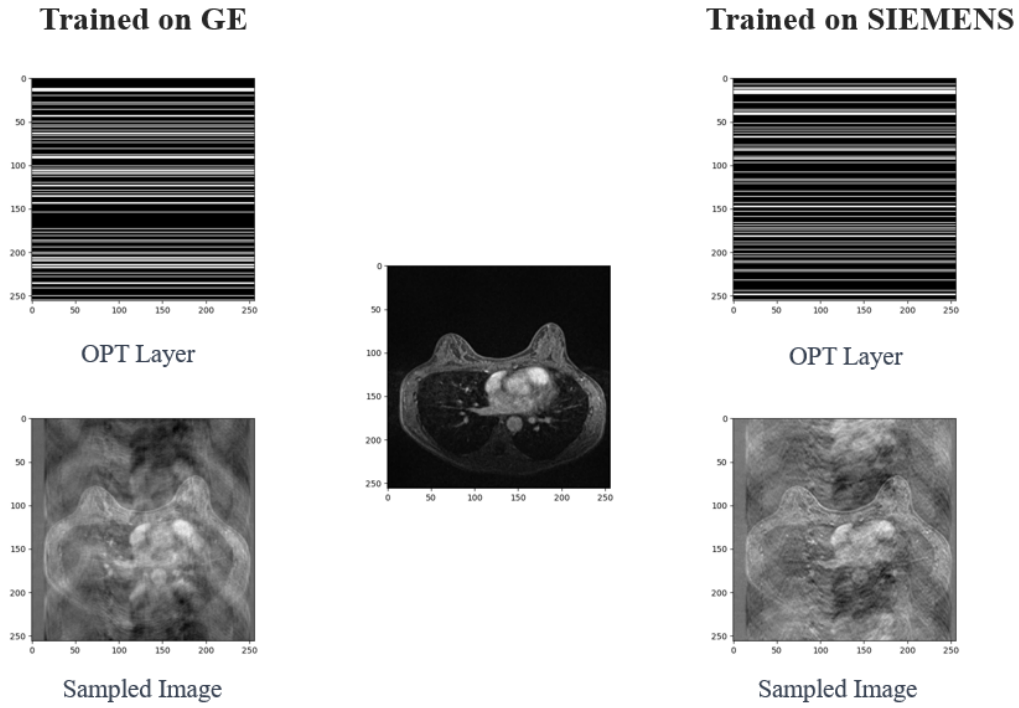

Figure 6: Optimized K-Sampling Layer: The left one is trained on GE, and the right one is trained on Simens. Sampled images' tumor regions are obvious

they are all greater than the AUC in the fully sampled cases. This demonstrates improved generalizability compared to using the full K-space data. Also, we can see that AUC is high when the sampling rate is low. In this case, we think it is because the limited features the model learns represent common features across scanners. Considering both robustness and generalizability, we chose 0.24 for GE-trained models and 0.22 for Siemens-trained models as the best K-sampling rates, as they provide the best balance between robustness and good generalizability. Future work could explore other sampling strategies or model architectures to further improve the performance and generalizability of the proposed method in medical image processing.

Table 1: AUC for different Sampling Rates at the End of Training

Trained on GE			Trained on Siemens		
Sampling Rate	Test on GE	Test on Siemens	Sampling Rate	Test on Siemens	Test on GE
0.148	0.837	0.689	0.113	0.822	0.726
0.242	0.878	0.625	0.222	0.868	0.714
0.355	0.884	0.621	0.308	0.859	0.703
0.496	0.882	0.605	0.464	0.862	0.731
0.601	0.889	0.605	0.578	0.879	0.734
0.664	0.898	0.635	0.674	0.873	0.748
0.710	0.897	0.594	0.750	0.855	0.743
0.804	0.902	0.630	0.781	0.881	0.743
0.886	0.887	0.609	0.882	0.854	0.756
1	0.890	0.659	1	0.873	0.750

5 Conclusion

In conclusion, this study has demonstrated that the performance of deep learning models can be impacted by variations in scanners and image quality levels. By employing a K-Sampling rate of approximately 0.2, we were able to achieve near-baseline accuracy for robustness. For generalizability, we observed high performance when using a sampling rate of 0.148 for training on GE and 0.882 for training on Siemens scanners.

Our approach involved the development of an optimized K-Sampling layer through automatic differentiation, effectively balancing both robustness and generalizability across different scanners. This finding is crucial for the application of deep learning models in medical image processing, as it highlights the importance of selecting an appropriate K-Sampling rate to maintain performance across diverse settings.

Future research could explore alternative sampling strategies or model architectures to further enhance the performance and generalizability of deep learning models in medical image processing. Additionally, investigating the underlying reasons behind the optimal performance at specific sampling rates could provide valuable insights for improving model robustness and generalizability. We hope that our findings will contribute to the ongoing advancement of deep learning techniques in medical imaging and foster further research in this area.

Acknowledgments

I would like to acknowledge Dr. Horstmeyer, Kanghyun Kim and Amey Chaware for their useful suggestions and guidance in this project.

References

[1] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[2] Gustav Mårtensson, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, Milica G Kramberger, et al. The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study. *Medical Image Analysis*, 66:101714, 2020.

[3] Xiaoqin Wang, Gongbo Liang, Yu Zhang, Hunter Blanton, Zachary Bessinger, and Nathan Jacobs. Inconsistent performance of deep learning models on mammogram classification. *Journal of the American College of Radiology*, 17(6):796–803, 2020.

[4] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

[5] Ashirbani Saha, Xiaozhi Yu, Dushyant Sahoo, and Maciej A Mazurowski. Effects of mri scanner parameters on breast cancer radiomics. *Expert systems with applications*, 87:384–391, 2017.

[6] Sreenath P Kyathanahally, André Döring, and Roland Kreis. Deep learning approaches for detection and removal of ghosting artifacts in mr spectroscopy. *Magnetic resonance in medicine*, 80(3):851–863, 2018.

[7] Wenjun Yan, Lu Huang, Liming Xia, Shengjia Gu, Fuhua Yan, Yuanyuan Wang, and Qian Tao. Mri manufacturer shift and adaptation: increasing the generalizability of deep learning segmentation for mr images acquired with different scanners. *Radiology: Artificial Intelligence*, 2(4):e190195, 2020.

[8] Ashirbani Saha, Michael R. Harowicz, Lars J. Grimm, Connie E. Kim, Sujata V. Ghate, Ruth Walsh, and Maciej A. Mazurowski. A machine learning approach to radiogenomics of breast cancer: A study of 922 subjects and 529 dce-mri features. *British Journal of Cancer*, 119(4):508–516, 2018.

[9] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.