Semantic-Aware Adaptive Binary Search for Hard-Label Black-Box Attack

A Thesis

Presented in Partial Fulfillment of the Requirements for the
Degree of Master of Science
with a

Major in Computer Science
in the
College of Graduate Studies
University of Idaho
by

Yiqing Ma

Approved by:

Major Professor: Min Xian, Ph.D.

Co-Major Professor: Aleksander Vakanski, Ph.D.

Committee Members: Frank Gao, Ph.D.

Department Administrator: Terence Soule, Ph.D.

May 2023

Abstract

Breast cancer is a major health concern globally, and early detection is crucial for successful treatment. Breast ultrasound is a widely used imaging modality for the diagnosis of breast cancer. In recent years, numerous studies have explored the use of deep learning for breast cancer classification in ultrasound images. These studies have shown promising results, with deep learning models achieving high levels of accuracy in detecting breast cancer. Despite the widely reported potential of deep neural networks for automated breast tumor classification and detection, these models are vulnerable to adversarial attacks, which can lead to significant performance degradation.

In this thesis, I build a novel adversarial attack approach under the decision-based black-box setting, where model details (e.g., architecture and parameters) are inaccessible, and querying the target model only provides the prediction of final class label (i.e., hard-label attack). The proposed attack approach has two major components: adaptive binary search and semantic-aware search. The adaptive binary search utilizes a coarse-to-fine strategy that applies different tolerance values in different searching stages to reduce unnecessary queries. The proposed semantic mask-aware search crops the search space by using breast anatomy, which significantly avoids invalid searches. The proposed approach is validated using a dataset of 3,378 breast ultrasound images and compared with other state-of-the-art methods by attacking three deep learning models. The results demonstrate that the proposed approach generates imperceptible adversarial samples at a high success rate (99.83%), and it dramatically reduces the average and median queries by 23.96% and 31.79%, respectively, compared with the state-of-the-art.

Acknowledgments

I would like to express my sincere gratitude to my advisor, Dr. Min Xian, for his invaluable guidance, mentorship, and patience throughout the entire journey of this thesis. His expertise, insight and feedback have been instrumental in shaping my research; guiding me through the writing process and making it the best it can be. I would like to thank my co-major professor Dr. Aleksandar Vakanski for his thoughtful contributions and constructive feedback throughout the research process. I would like to thank my committee member Dr. Frank Gao. I wanted to express my gratitude for the crucial research skills that I learned from him. The knowledge and tools that I gained through you're his instruction have been instrumental in my academic and professional growth. I am deeply grateful for their dedication to my success and for the countless hours they have invested in helping me reach this milestone. Thank you for believing in me, challenging me, and supporting me in every step of this journey. Finally, I would also like to thank the faculty and staff in Department of Computer Science at the University of Idaho.

Dedication

Most importantly, I would like to thank my family. To my beloved parents, who have provided me with a high-quality education and encouraged me throughout my academic journey. Their unwavering love and understanding have been my source of strength and inspiration. Without their tireless support and innumerable sacrifice, it would be hard for me to have made this achievement possible. I am truly thankful and honored to have them as my parents and I will always cherish the role they have played in my life. I would like to thank all of the participants in my study for their willingness to take part and share their experiences with me. Without their contributions, this research would not have been possible.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Research Problem

Breast cancer has emerged as one of the most prevalent types of cancer globally, contributing to nearly 12% of all newly diagnosed cancer cases (American Cancer Society). Breast cancer is estimated to affect around 31% of female cancer cases in U.S. in 2023 (American Cancer Society). Although Deep Neural Networks (DNNs) demonstrated unprecedented performance in medical image classification, recent research (Szegedy, Zaremba and Sutskever; Goodfellow, Shlens and Szegedy) indicated that DNNs as well as conventional machine learning (ML) models can be compromised by adversarial samples. That is, adversarial samples can be synthesized by adding unnoticeable perturbations to clean inputs and cause the target DNNs to misclassify such samples and return incorrect class labels. Adversarial attacks (Madry, Makelov and Schmidt; Carlini and Wagner; Moosavi-Dezfooli, Fawzi and Frossard) have been realized to achieve high attack success rates by introducing low levels of perturbations.

A predominant portion of existing adversarial attacks was designed for evading ML models for the classification of natural image datasets (e.g., CIFAR-10 (Krizhevsky) and ImageNet (Deng, Dong and Socher)). A body of works demonstrated black-box attacks on medical images as well. However, medical images possess domain-specific characteristics distinct from natural images. As a result, black-box evasion attacks are less successful with some modalities of medical images in terms of the number of model queries and success rate. To overcome the challenges in existing attacks, I propose an adaptive binary search and semantic mask-aware search to reduce the number of queries in extreme examples through coarse search and refined search. The thesis is motivated by declining the swing search of the decision boundary at an early stage and narrowing down search regions in consonance with search depth. Comprehensive experiments are conducted using a combination of several breast ultrasound datasets (Al-Dhabyani, Gomaa and Khaled; Shareef, Xian and Sun; Yap, Pons and Marti; Geertsma) and demonstrate performance of state-of-the-art models.

Adversarial attack research focuses on studying and understanding the vulnerabilities of ML models to malicious attacks. The purpose of this research is to identify weaknesses in ML

models and to develop techniques to defend against adversarial attacks to ensure the reliability and trustworthiness of ML-based systems.

1.2 Thesis Objectives and Contributions

The thesis is organized as follows: Chapter 2 is the literature review which discusses existing research on different adversarial attack categories and the breast ultrasound image classification task. Chapter 3 presents the proposed methodology. It describes two major designs of this thesis: adaptive binary search and semantic-aware search. Chapter 4 describes four datasets that will be used for experiments, and the data preprocessing steps. Chapter 5 depicts the testing environment and four different experiments based on the proposed method.

The primary contributions of the study are summarized below.

- The proposed adaptive binary search algorithm effectively reduces unnecessary queries by searching adversaries using a coarse-to-fine manner.
- The proposed semantic-aware search algorithm avoids invalid searches by cropping the search space using semantic masks from breast anatomy.
- The combination of the two above algorithms leads to a novel hard-label black-box attack approach. It significantly reduces the number of queries for searching adversaries for extreme samples.

Chapter 2: Related Work

2.1 Attack Categories

Adversarial attacks can be categorized into white-box attacks and black-box attacks. In a white-box adversarial setting, attackers are assumed to have complete knowledge of the targeted model, including knowledge of the model architecture, parameters, gradients, objective function, etc. The black-box setting is more challenging because adversaries do not have access to the model structure or parameters. It is also more realistic since most model developers do not provide such access to users. In both white-box and black-box attacks, adversarial attacks can be identified as targeted and untargeted according to the final output of the attack result. A targeted attack fools a model into falsely predicting a specific label for the adversarial image. An untargeted attack classifies predicted irrelevant label of the adversarial, which means it is not the correct label. Adversarial attacks have two scenarios evasion attack and poison attack. The attacker injects fake training data intending to corrupt the learned model in a poisoning attack. Only malicious samples in the test set are modified to evade detection and misclassified as legitimate in an evasion attack, thus evasion attack does not influence the training data. This thesis mainly focuses on evasion attacks under the black-box setting.

2.2 White-box Attack

L-BFGS (Szegedy, Zaremba and Sutskever) was the first approach that demonstrated that adding a small perturbation to an image could evade a target ML classifier and produce the wrong classification. FGSM attack (Goodfellow, Shlens and Szegedy) used the sign of the gradient of a neural network loss with respect to inputs to find adversarial perturbations. The sign function is used to ensure that the perturbation is in the direction that maximizes the loss function. PGD attack (Madry, Makelov and Schmidt) applied FGSM iteratively with a smaller distortion in each step to minimize the overall perturbation. CW attack (Carlini and Wagner) formulated the adversarial example generation as a constrained optimization problem, that approximates the minimal perturbation for misclassifying an input sample.

2.3 Black-box Attack

In real-world scenarios, the model structure will not be provided to the attacker. Therefore, the adversary has limited knowledge of the model. The attacker can apply a black-box attack to a model with similar architecture, then transfer to a targeted model. This is called transfer-based black-box attacks. Without employing any extra model, the adversarial black-box attack can be categorized by model feedback from given queries. The black-box attack is split into scored-based attacks and decision-based attacks.

2.4 Transfer-based Black-box Attack

Prior works (Papernot, McDaniel and Goodfellow) demonstrated that adversarial samples from one ML model could transfer to other models in a black-box setting. Transfer-based black-box attacks are black-box attacks that steal information from a surrogate model. The surrogate model is considered a white-box model and the attacker can implement any white-box techniques to that. Guessing smart (Brunner, Diehl and Le) computed gradients from surrogate models and find an adversarial position, which is helpful to move a small distance toward a clean image.

2.4.1 Score-based Black-box Attack

Scored-based black-box attack (also called soft-label attack) obtains information about a target ML model by querying the model. In a scored-based attack, an adversary can acquire probability confidence scores from the targeted model. Zeroth order optimization approaches (Chen, Zhang and Sharma) employed returned confidence scores to estimate the gradient and generate adversarial samples. Also, SimBA (Guo, Gardner and You) purposed a low-frequency perturbation, a random direction is repeatedly selected from orthogonal search directions, using confidence scores to check whether it is pointing toward or away from the decision boundary. SignHuner (Al-Dujaili and O'Reilly) estimated the gradient sign bits (Bernstein, Wang and Azizzadenesheli) based on the gradient from the previous step to reach faster convergence. Square attack (Andriushchenko, Croce and Flammarion) used a score-based setting without relying on local gradient information by utilizing a randomized search at each step to find an optimized solution.

2.4.2 Decision-based Black-box Attack

Decision-based attack (also called hard-label attack) returns only the top-1 class prediction from the target model. Boundary attack (Brendel, Rauber and Bethge) indicated decision-based attack, which required less hyperparameter than transfer-based attack and was more robust than score-based attack. The decision-based black-box attack is closer to realistic scenarios. HJSA (Chen, Jordan and Wainwright) utilized binary information at the decision boundary to calculate the directional gradient. SignOPT (Cheng, Singh and Chen) extended the previous work OPT (Cheng, Le and Chen), and specified a single query oracle for computing zeroth-order gradient direction. RayS (Chen and Gu) used the research in (Al-Dujaili and O'Reilly; Cheng, Singh and Chen), and implemented a hyperparameter-free decision-based attack without zeroth-order gradient estimation. It significantly reduced the number of queries for attacking DNNs to several hundred. But it failed to use a small number of queries to find adversaries for many extreme BUS images.

2.5 Breast Ultrasound Image (BUS) Classification

Recent studies showed DNN enhanced the classification of breast ultrasound images. Hijab et al. (Hijab, Rushdi and Gomaa) indicated deep learning outcomes in biomedical applications could be significantly enhanced through the development of pre-trained and fine-tuned convolutional neural network (CNN) architecture using medical imaging data. Xie et al. (Xie, Song and Zhang) designed a dual-sampling convolutional neural network (DSCNN) with residual networks for the diagnosis of breast tumor images. Their network could prevent gradient disappearance and degradation and improve accuracy by using a parallel DSCNN. ESTAN (Shareef, Vakanski and Freer) presented a new architecture that utilizes two encoders to extract global and local information and integrate image context information at varying scales. The authors introduced row-column-wise kernels that conform to the horizontal arrangement of the breast anatomy tissue layers in BUS images. By employing this approach, their network demonstrated enhanced segmentation performance for tumors of varying sizes and surpassed the performance of existing state-of-the-art methods for BUS segmentation. MT-ESTAN (Shareef, Xian and Sun) conducted a multi-task neural network that combines two separate tasks tumor classification (primary task) and tumor segmentation (secondary task) associated with their previous work ESTAN in one shared model. The network learns from both tasks and alleviates the low generalization issue

caused by small training datasets. The learned shared features between object segmentation and classification improve the robustness and generalizability of the model.

However, Ma et al. (Ma, Niu and Gu) implemented FGSM, BIM (Kurakin, Goodfellow and Bengio), PGD and CW on medical images and proved medical deep learning systems are vulnerable to small carefully-engineered perturbations. They explained the reasons are the complex biological textures in medical images that cause higher gradient regions, and those regions are sensitive to small adversarial perturbations. Moreover, DNNs that are currently considered for the natural image at a large scale, the model may not be optimized for medical imaging tasks due to overparameterization. This can lead to a loss landscape that is too sharp and an increased susceptibility to adversarial attacks.

Chapter 3: Proposed Method

3.1 Problem Formulation (Preliminary)

Let f be a DNN model, x_0 is the input (clean image) and y is the true class label associated with x_0. A general hard-label black-box adversarial attack can be divided into two categories: targeted attack and untargeted attack, which are formulated as

$$\text{Targeted Attack: } minimize\ D(x, x_0)\ such\ that\ f(x) = t \quad (1)$$

$$\text{Untargeted Attack: } minimize\ D(x, x_0)\ such\ that\ f(x) \neq y \quad (2)$$

The goal is to generate an adversarial sample x that is nearest to the x_0, D denotes the distance between x and x_0 under e.g., the L_∞ norm. The targeted attack is to change the decision to some pre-specified label t. $D(x, x_0) < \epsilon$, where ϵ is the maximum perturbation strength under L_∞ norm. This thesis only considers untargeted attacks Eq. (2). Optimizing this framework under soft-label black-box attack can use the logits from a surrogate model or probabilities of top-k predictions (Moosavi-Dezfooli, Fawzi and Frossard; Moon, An and Song; Andriushchenko, Croce and Flammarion; Chen, Zhang and Sharma; Guo, Gardner and You). However, the hard-label black-box is more challenging since the model only provides the top-1 classification result. (Cheng, Le and Chen; Cheng, Singh and Chen) re-formulated hard-label black-box attack by finding the direction θ with the shortest distance to the decision boundary.

$$minimize\ g(\theta) \quad (3)$$

$$g(\theta) = minmize\ \lambda \quad s.t.\ f\left(x_0 + \lambda \frac{\theta}{||\theta||}\right) \neq y \quad (4)$$

In E.q. (4), $g(\theta)$ is the distance from x_0 to closest adversarial sample along the search direction θ; λ is the distance to the decision boundary which is determined by a binary search algorithm.

RayS (Chen and Gu) significantly improved the optimization of the above-formulated problem by bounding the search space to a discrete set of ray directions, i.e., $\theta \in [-1, 1]^d$ where d is the number of dimensions of the space. It also modified the optimization

framework E.q. (4), to a none zeroth-order gradient estimation on L_∞ norm ball. RayS reduced the possible searching directions from R^d (Cheng, Le and Chen; Cheng, Singh and Chen) to 2^d . RayS used a greedy search algorithm without estimating any gradients, which only stored the best search direction established on the previous search direction. Different from (Cheng, Le and Chen; Cheng, Singh and Chen; Chen, Jordan and Wainwright), taking multiple or an average of few queries to determine the next search direction, RayS queried $f(x)$ once before doing a binary search for λ. The search direction is selected to do a binary search, only if an adversary sample x could be found. Nevertheless, RayS lacked a reliable way to look for an adversarial example when block partitions rise. The reason is the search directions get closer to each other. Their small constant binary search tolerance causes more unnecessary queries from hierarchical search.

The proposed method is a novel approach that applies adaptive binary search and semantic-aware search to reduce the search queries on extreme examples. The proposed method followed RayS equation (4).

3.2 Adaptive Binary Search

The binary search has been widely adopted in hard-label black-box attacks (Chen, Jordan and Wainwright; Cheng, Le and Chen; Cheng, Singh and Chen; Chen and Gu) to efficiently search adversaries along a direction Eq. (3). It attempts to minimize the distortion to generate adversarial examples close to the decision boundary. The efficiency of the binary search depends on the size of the search interval and the number of queries required to locate the target value. The number of queries in binary search is directly determined by a tolerance value that refers to the acceptable precision of the target adversary. For a specific search direction in the binary search algorithm, an adversary should meet the following condition defined by tolerance.

$$\left|\left|x_e - x_s\right|\right|_\infty \leq \tau \; s.t. f(x_s) = y \; and \; f(x_e) \neq y \qquad (5)$$

where x_s and x_e are two samples on the two sides of the decision boundary. τ was defined as a fixed small positive value in previous work (Chen and Gu).

A small τ allows find adversaries close to a decision boundary but needs more queries to search for a valid match. On the other hand, a large tolerance requires fewer queries but

generates adversaries that are not close to a decision boundary. RayS (Chen and Gu) used a hierarchical searching strategy and a fixed small τ to search a sequence of adversaries to approach the searching goal defined by the distance upper bound ϵ. Because of the fixed small τ , all adversaries are searched using the same precision, resulting in many unnecessary queries. Especially for a deep search hierarchy, this may lead to an enormous number of queries. Inspired by the trade-off between tolerance values and the number of queries, this thesis proposes an adaptive binary search (AdptBS) in Algorithm 1 which applies a larger τ in the coarse search stage and a small τ at a fine search stage. As shown in line 9 of Algorithm 1, the fine search stage is defined by a distance range $[\epsilon + \mu * \tau, \epsilon + \tau]$ which uses small τ . For the coarse search stage, τ is set to a large value (0.1) that needs fewer queries. μ is a parameter to adjusted when to change τ. This control τ is always less than the difference between $D(x, x_0)$ and distance upper boundary ϵ. Algorithm 1 adjusted τ, if current adversarial sample is in the blue region in Figure 1.

Algorithm 1 AdptBS

Input: Model f, clean image x_0, label y, distance upper bound ϵ, search direction θ, best radius r_{best}, binary search tolerance τ, tolerance adjust lower boundary μ

1 $\theta_u = \frac{\theta}{||\theta||_2}$
2 **if** $f(x_0 + r_{best} \cdot \theta_u) == y$ **then** ▷ Invalid search direction
3 return τ, ∞
4 $x_s = x_0$, $x_e = x_0 + min(r_{best}, ||\theta||_2) \cdot \theta_u$
5 **if** $\epsilon + \mu * \tau < \mathrm{D}(x_e, x_0)_\infty < \tau + \epsilon$ **then**
6 $\tau \mathrel{/}= 10$
7 **while** $||x_e - x_s||_\infty > \tau$ **do**
8 $x_m = (x_s + x_e)/2$
9 **if** $f(x_m) \neq y$ **then**
10 $x_e = x_m$
11 **else**
12 $x_s = x_m$
13 return τ, $||x_e - x_0||_\infty$

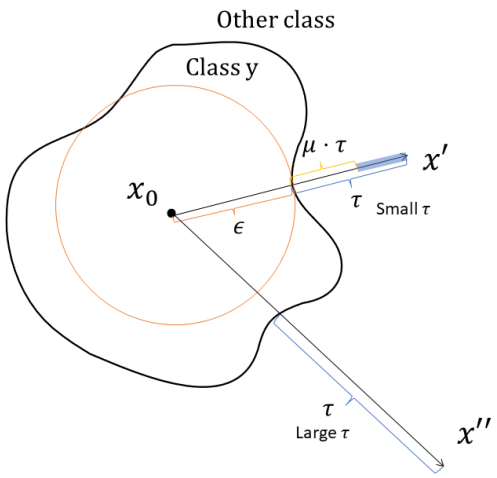

Figure 1. Adaptive tolerance range in Algorithm 1.

3.3 Semantic-Aware Search

RayS (Chen and Gu) searches adversarial samples using the whole image space with high dimensions. The worst case of RayS strategy is searching for each pixel on the whole image. Guessing smart (Brunner, Diehl and Le) elucidated regional masking for an attack to limit the perturbation to specific regions to reduce the dimensionality of the search space to avoid waste of queries. Thus, they designed an attack to take larger steps in high-difference regions. Guessing smart directly reduces the searching space into mask region with the current position. MT-ESTAN (Shareef, Xian and Sun) improved the breast ultrasound cancer classification with a multiclass learning approach, joint breast classification and tumor segmentation. In this work, semantic masks of BUS images are used to choose image regions of breast tumors and mammary tissues, shown in Figure 2. The regions are selected because tumor classes are mainly determined by image features in the regions, and it is more efficient to search adversaries by adding perturbations to these regions.

Inspired by the regional masking approach (Brunner, Diehl and Le; Shareef, Xian and Sun), this work proposes a hybrid strategy to produce a search mask for each BUS image by combining a semantic mask and a cropping mask. The semantic mask is generated by using a U-Net-based semantic segmentation approach; and the cropping mask trims blocks at the front and end of the search direction. As shown in Algorithm 2. The algorithm skips the front

and end blocks using $crop_k$ of the search direction to ignore top and bottom image regions. In Figure 4, the cropping mask can be considered as a pruned branch to remove the unnecessary search space compared with RayS (Figure 3). The combination of semantic masks and cropping masks skips checking a region that does not contain crucial features. A block-level cut point K is applied to increase the skipping blocks to adapt to the increasing number of fine blocks generated by the $BlockSplit(k)$ function. The details about the best cut point K will be discuss in Chapter 5.

Algorithm 2 Semantic AdptBS

Input: : Model f, clean image x_0, label y, distance upper bound ϵ, semantic mask M

Initialize search direction $\theta_{best} = (1, \cdots, 1)$, radius $r_{best} = \infty$, and block level $k = 0$
Initialize binary search tolerance $\tau = 0.1$, block level cut point K, cropping block size $crop_k$
function BLOCKSPLIT(k)
    cut θ_{best} into 2^k blocks of equal size and save the splitting blocks into a list
    return Blocks list
while remaining queries > 0 **do**
    $\theta_{tmp} = \theta_{best}$
    blocks = BlockSplit(k)
    if $k > K$ **then**
        $crop_k$ *= 2 ▷ skip more blocks for fine splitting
    for i in blocks **do**
        $\theta_{tmp}[i]$ *= -1
        if $i \in$ ((blocks[$crop_k$:$-crop_k$] $\cup$ M) **then**
            τ, r_{tmp} = AdptBS($f, x_0, y, r_{best}, \theta_{tmp}, \epsilon, \tau$)
            if $r_{tmp} < r_{best}$ **then**
                $r_{best} = r_{tmp}$, $\theta_{best} = \theta_{tmp}$
    k += 1
return $x_0 + \theta_{best} * r_{best}$

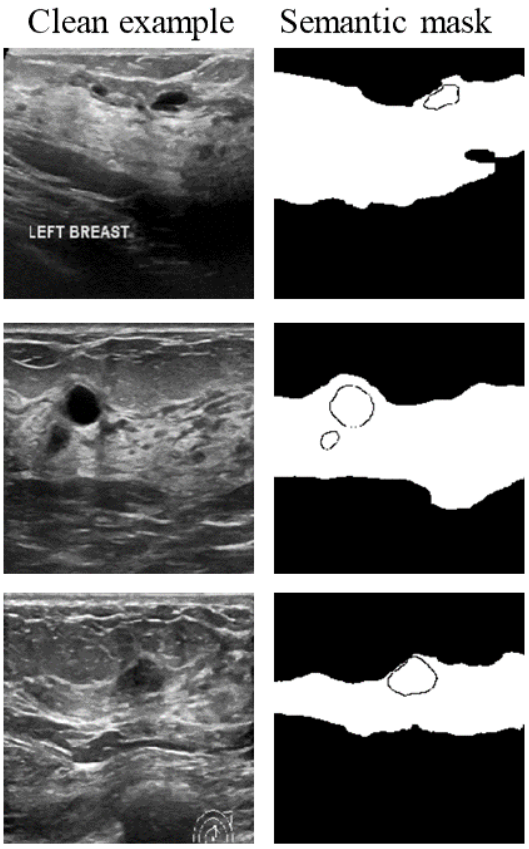

Figure 2. Semantic masks of BUS images.

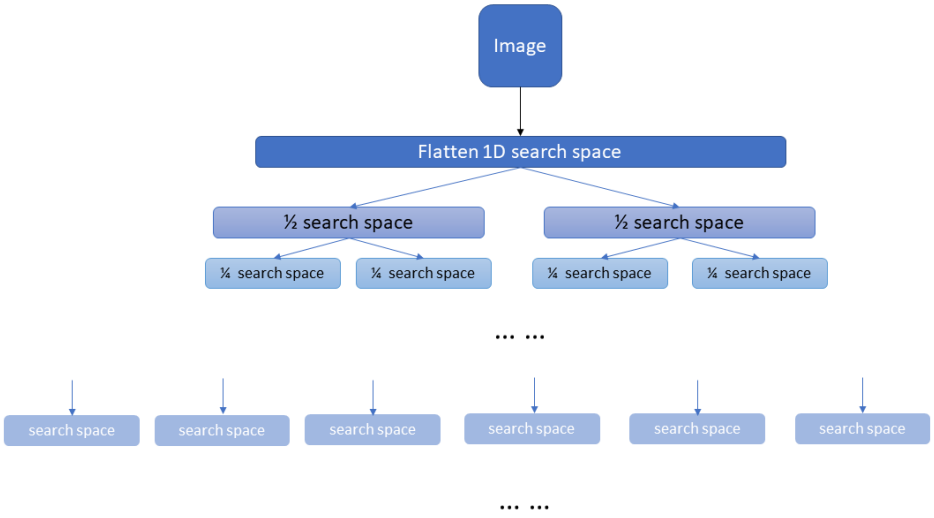

Figure 3. RayS block splitting.

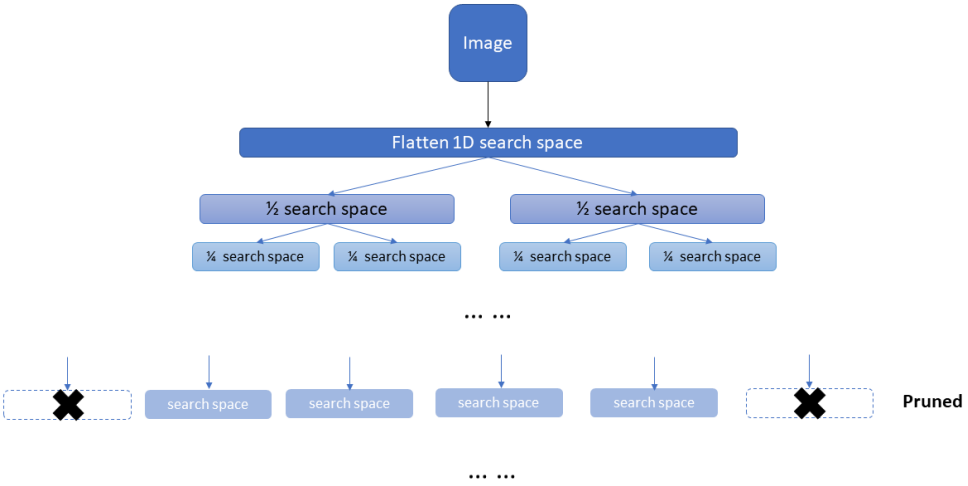

Figure 4. Block splitting in semantic-aware search.

Chapter 4: Materials and Data

4.1 Breast Ultrasound Dataset

The method was evaluated on Breast Ultrasound (BUS) dataset. Since the existing public BUS Datasets are small, the performance of all attack algorithms was tested on a combination of four BUS datasets: BUSI (Al-Dhabyani, Gomaa and Khaled), BUSIS (Zhang, Xian and Cheng), HMSS (Geertsma) and Dataset B (Yap, Pons and Marti). Since BUS problem is binary classification, benign and malignant cases are considered. The dataset contains a total of 3378 images: 1698 benign images and 1680 malignant images. BUSI dataset contains a total of 647 images with 437 benign images and 210 malignant images. The images in BUSI were acquired by LOGIQ E9 ultrasound and LOGIQ E9 Agile ultrasound system. These images were from Baheya Hospital for Early Detection & Treatment of Women's Cancer in Cairo, Egypt. BUSIS dataset contains a total of 562 images with 306 benign and 256 malignant images from China. The images were collected by multiple ultrasound devices: GE VIVID 7, LOGIQ E9, Hitachi EUB-6500, Philips iU22 and Siemens ACUSON S2000. The dataset was prepared by the Second Affiliated Hospital of Harbin Medical University, the Affiliated Hospital of Qingdao University and the Second Hospital of Hebei Medical University. HMSS dataset contains a total of 2006 images with 846 benign and 1160 malignant images. It was collected by Dr. Geertsma at Gelederse Vallei hospital in Netherlands, in a collaboration with Hitachi Medical Systems Europe. Dataset B contains a total of 163 images with 109 benign and 54 malignant images from different women in 2012. It was provided by UDIAT Diagnostic Centre of the Parc Taulí Corporation, Sabadell (Spain) and images were gathered by a Siemens ACUSON Sequoia C512 system 17L5 HD linear array transducer (8.5 MHz).

4.2 Data Preprocessing

Most images in BUS datasets are rectangular. Because Deep Neural Network models need a square input shape, images are modified to the square shape. If reshaped images from a rectangular shape to a square shape will bring out the distortion problem of the tumor region. To avoid morphologic changes in the breast tumors and tissue regions, all images are zero-padding and fulfilled to form a square following MT-ESTAN (Shareef, Xian and Sun) setting.

After padding all images, input images and corresponding semantic masks are reshaped to a fixed size 224×224.

Chapter 5: Experimental Results

5.1 Targeted Model Settings

All experiments use a randomly selected set of 700 images from the dataset for testing, and use rest of the images for training. All experiments use pretrained ResNet50 (He, Zhang and Ren), VGG16 (Simonyan and Zisserman) and DenseNet121 (Huang, Liu and Maaten), and their accuracy is 82.94%, 80.7% and 87.83%, respectively. All three models have been trained 100 epochs with the Adam optimizer, learning rate of 0.0001, and batch size of 4. ResNet50 is used to compare the effectiveness of adaptive binary search, semantic-aware search and overall performance on other hard-label black-box attacks. The best parameters from adaptive binary search and sematic-aware search will be validated on VGG16 and DenseNet121, aiming to illustrate the approach is aggressive to deep neural networks.

5.2 Adversarial attack settings

Following the adversarial attack setting in RayS (Chen and Gu), the distance upper bound ϵ is set to 0.05, and the maximum number of queries is 10,000 for all attacks. The criterion for a successful attack is the L_∞ between the adversarial sample and clean image is less than the pre-defined ϵ.

5.3 Experiment Environments and Evaluation Metrics

All neural network models and adversarial attacks are implemented using Python 3.7.0, Keras 2.3.1, TensorFlow 1.13.1 and Pytorch 1.9.1. All experiments were conducted with NIVIDA Quadro RTX 8000 GPUs, equipped with CUDA Toolkit 10.2. The number of average (AVG) and median (MED) queries and the attack success rate (SR) are used to quantitatively evaluate the performance of different adversarial attacks. The number of average and median queries is counted from successfully attacked images on the test set. The success rate is a ratio between successful attacks and total attacks, and it is calculated only from images that were correctly classified by the targeted model on the test set.

5.4 The Effectiveness of Adaptive Binary Search

In this section, different binary search strategies in the RayS attack (Chen and Gu) framework are validated using ResNet50. The original RayS set a fixed binary search tolerance τ to 0.001. Results of RayS with other τ values and AdptBS are reported in Table 1.

The results in the first three rows show that the number of queries (AVG and MED) is sensitive to τ ; and small τ need more queries to find adversaries and large τ can significantly reduce the queries. But large τ could also have the risk to decrease the success rate. However, the attacker cannot find out the best τ without several attempts and multiple attacks on the model, which is not efficient. Thus, an adaptive τ can automatically obtain the best value for searching the initial attack starting point and adjust tolerance based on the current distance between adversarial samples and the decision boundary. The proposed AdaptBS is applied to replace the binary search algorithm in the RayS attack, and its results are shown in the last three rows of Table 1. AdptBS halves the number of queries without affecting attack success rate. The AdptBS method preserves the high success rate (99.83%) of the original RayS with fixed small τ , and reduces the AVG and MED queries by 21.47% and 29.37% respectively. μ decides when to change τ, the best result is $\mu = 0.9$ and the performance reduce when using smaller μ. $\mu = 0.9$ restricts τ quickly narrow to a very small value, which use large τ for immense perturbation and use small τ when perturbation in close proximity to ϵ.

Table 1. Results of attack with different binary search methods. The percentage values in the parenthesis show the difference between the baseline model and a new design.

Attack Method	μ	τ	Queries(AVG)↓	Queries(MED) ↓	SR(%) ↑
RayS (Chen and Gu)	-	0.001(original)	411.94	248.5	99.83
	-	0.1	299.06 (-27.40%)	159.0 (-36.01%)	99.15 (-0.68%)
	-	0.01	346.07 (-15.99%)	206.0 (-20.63%)	99.83
AdptBS	0.9	Adaptive	323.47 (-21.47%)	175.5 (-29.37%)	99.83
	0.8	Adaptive	325.91 (-20.88%)	178.5 (-28.16%)	99.83
	0.7	Adaptive	330.26 (-19.82%)	181.0 (-27.16%)	99.83

5.5 The Effectiveness of Semantic-Aware Search

In this section, semantic-aware search in the RayS and semantic-aware search are validated using ResNet50. The semantic-aware search aims to reduce the search space. It is integrated into RayS and the proposed AdptBS algorithm and the results are demonstrated in Table 2.

The original RayS with the proposed semantic-aware search reduces AVG, and MED queries by 3.87% and 4.22%, respectively. The parameters are $\tau = 0.001$, $crop_k = 2$ and $K = 5$. The semantic-aware search with the proposed AdptBS algorithm can dramatically drop the AVG queries by 23.96%, and the MED queries by 31.79% with $\mu = 0.9$, $crop_k = 2$ and $K = 5$. The impressive results demonstrate adding small perturbations only to breast tumors and mammary regions could find good adversaries more efficiently. The best $crop_k$ is 2, Table 2 illustrates SR reduce when $crop_k$ increased to large value. The purpose method drops less search directions at the beginning of the search stage. Otherwise, the purpose method throws more blocks when splitting search space into small-scale search directions, because the neighbor search directions are close to each other. The average of split blocks in RayS are distributed at 5 and 6. Table 2 shows MED queries in Semantic-Aware AdptBS for $K = 5$ and $K = 6$ are both 169.5. Since $K = 5$ performed better on AVG queries, the rest the experiments are all setting $K = 5$. The purpose method is also effectively reducing the number of queries for searching adversaries for extreme samples. There are 15.68% of test images use larger than 400 queries and 9.61% of test images use larger than 600 queries in the purpose approach. However, 29.51% of test images use more than 400 queries and 15% of test images use larger than 600 queries in RayS. Instances of extreme cases on both approaches are shown in Figure 5.

Table 2. Results of attack with semantic-aware search.

Attack Method	μ	$crop_k$	K	Queries(AVG)↓	Queries(MED) ↓	SR(%) ↑
RayS (Chen and Gu)	-	2	5	411.94	248.5	99.83
RayS + Semantic Mask	-	2	5	395.98 (-3.87%)	238.0 (-4.22%)	99.83
	-	2	4	402.29 (-0.23%)	236.5 (-4.82%)	99.83
	-	2	6	399.68 (-2.97%)	238.5 (-4.02%)	99.83
	-	3	5	404.70 (-1.75%)	243.5 (-2.01%)	99.83
	-	4	5	402.29 (-2.34%)	236.5 (-5.07%)	99.83
	0.9	2	5	313.22 (-23.96%)	169.5 (-31.79%)	99.83
	0.9	2	4	308.62 (-25.07%)	164 (-34.00%)	99.66 (-0.17%)

Semantic-Aware AdptBS	0.9	2	6	318.21 (-22.75%)	169.5 (-31.79%)	99.83
	0.9	3	5	309.18 (-24.94%)	170.0 (-31.58%)	99.49 (- 0.34%)
	0.9	4	5	308.62 (-25.08%)	164.0 (-34.00%)	99.66 (-0.17%)
	0.7	2	5	318.42 (-22.70%)	175.5 (-29.37%)	99.83
	0.7	3	5	328.57 (-20.23%)	175.0 (-29.57%)	99.83
	0.8	2	5	314.04 (-23.76%)	173 (-30.38%)	99.83
	0.8	3	5	319.94 (-22.33%)	173 (-30.38%)	99.66 (-0.17%)

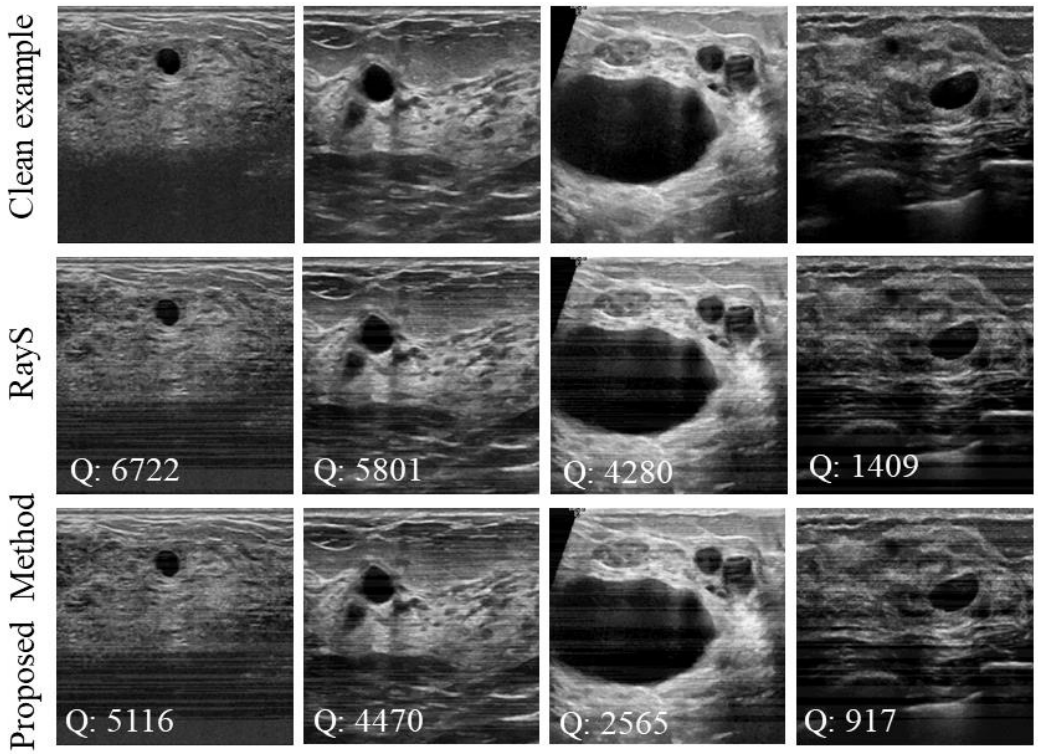

Figure 5. Results of RayS and purpose method on extreme images. Q: number of queries.

5.6 Attack on Other Deep Classifiers

RayS and the proposed method are used to attack three deep learning models, ResNet50, DenseNet121, and VGG16. The three models are pretrained and finetuned on BUS images training set. As shown in Table 3, the proposed method outperforms RayS in terms of AVG and MED queries on attacking all three models. The tolerance τ of RayS is 0.001 for all

three target models. The parameters of Semantic-Aware AdptBS for three target models are $\mu = 0.9$, $crop_k = 2$ and $K = 5$. For VGG16, the MED queries of the proposed method are 33.67% less than that of the RayS. For Desenet121, the MED queries of the proposed method are 32.03% less than that of the RayS. In addition, Densenet121 required more queries during attacks, which indicates that the model is more robust than the ResNet50 and VGG16.

Table 3. Results on attacking three models.

	Model	Queries(AVG)↓	Queries(MED) ↓	SR(%) ↑
RayS (Chen and Gu)	ResNet50	411.94	248.5	99.83
	Desenet121	618.67	384.0	99.52
	VGG16	456.06	297.0	100
Semantic-Aware AdptBS	ResNet50	313.22	169.5	99.83
	Desenet121	509.28	261.0	99.52
	VGG16	368.68	197.0	100

5.7 Compared with the State-of-the-art Attacks

Three state-of-the-art hard-label black-box attack approaches (i.e., OPT (Cheng, Le and Chen), SignOPT (Cheng, Singh and Chen), and RayS) are compared with the proposed method. The three attacks use the fixed tolerance for binary search. ResNet50 is used as the baseline classifier. As shown in Table 4, Sign-OPT and OPT randomly initialize a starting point with Gaussian noise or uniform noise and require many queries when using binary search to find the direction with the shortest distance to the decision h boundary and calculate the directional derivative. Each search direction generated by random noise needs a binary search to find the closest distance to the decision boundary. The binary search for each direction causes a massive number of queries. RayS significantly improves the success rate and reduces the queries to only several hundred due to its new search strategy in a discrete space. These three attacks all use the same tolerance for binary search, which prompts more queries to find an adversarial sample close to the decision boundary. Moreover, all three

attacks search the entire image for each iteration, which is a large search space. Local semantic-aware search shrinks search space to significant features. The proposed method achieves the same SR as RayS but outperforms the other three approaches on AVG and MED queries.

Table 4. Performance of the state-of-the-art hard-label black-box attack approaches.

Method	Queries(AVG)↓	Queries(MED) ↓	SR(%) ↑
OPT (Cheng, Le and Chen)	3218.36	2120.5	33.72
Sign-OPT (Cheng, Singh and Chen)	7066.05	7137.0	24.78
RayS (Chen and Gu)	411.94	248.5	99.83
Semantic-Aware AdptBS	313.22	169.5	99.83

Chapter 6: Conclusion

This thesis introduced a novel back-box adversarial attack approach against deep learning classifiers for breast ultrasound images. It only requires hard-label predicted outputs by the target model for the generation of adversarial samples. The proposed attack method integrates the semantic-aware search and adaptive binary search and outperforms state-of-the-art approaches in terms of average queries and success rate. The adaptive binary search component allows selecting an accommodative tolerance for binary search in different search stages. Using a semantic mask reduces the attack search space, which is critical due to the tremendous impact on model prediction. Experimental results on a large dataset of breast ultrasound images demonstrated the query-efficiency and the effectiveness of the proposed back-box attack.

References

Al-Dhabyani, Walid, et al. "Dataset of breast ultrasound images." *Data in brief* Vol.28 104863 (2019).

Al-Dujaili, Abdullah and Una-May O'Reilly. "Sign Bits Are All You Need for Black-Box Attacks." *International Conference on Learning Representations*. 2020. https://openreview.net/forum?id=SygW0TEFwH.

American Cancer Society. 2022. Cancer Facts & Figures. https://cancerstatisticscenter.cancer.org/#!/.

Andriushchenko, Maksym, et al. "Square Attack: a query-efficient black-box adversarial attack via random search." *arXiv: 1912.00049 [cs.LG]*. 2020.

Bernstein, Jeremy, et al. "signSGD: Compressed optimisation for non-convex problems." *International Conference on Machine Learning*. PMLR, 2018. 560--569.

Brendel, Wieland, Jonas Rauber and Matthias Bethge. "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models." *ArXiv* abs/1712.04248 (2017).

Brunner, Thomas, et al. "Guessing smart: Biased sampling for efficient black-box adversarial attacks." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019. 4958--4966.

Carlini, Nicholas and David Wagner. "Towards evaluating the robustness of neural networks." *2017 ieee symposium on security and privacy (sp)*. Ieee, 2017. 39--57.

Chen, Jianbo, Michael I. Jordan and Martin J. Wainwright. "HopSkipJumpAttack: A Query-Efficient Decision-Based Attack." *2020 IEEE Symposium on Security and Privacy (SP)*. 2020. 1277-1294.

Chen, Jinghui and Quanquan Gu. "RayS: A Ray Searching Method for Hard-label Adversarial Attack." *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2020).

Chen, Pin-Yu, et al. "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models." *In Proceedings of the 10th ACM workshop on artificial intelligence and security*. 2017. 15-26.

Cheng, Minhao, et al. "Query-Efficient Hard-label Black-box Attack:An Optimization-based Approach." *arXiv:1807.04457 [cs.LG]* (2018).

—. "Sign-OPT: A Query-Efficient Hard-label Adversarial Attack." *arXiv:1909.10773v3 [cs.LG]* (2020).

Deng, Jia, et al. "ImageNet: A large-scale hierarchical image database." *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009. 248-255.

Geertsma, Taco. *Ultrasoundcases.info*. 2014. FujiFilm. <https://www.ultrasoundcases.info/>.

Goodfellow, Ian J., Jonathon Shlens and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Guo, Chuan, et al. "Simple black-box adversarial attacks." *In International Conference on Machine Learning*. PMLR, 2019. 2484-2493.

He, Kaiming, et al. "Deep Residual Learning for Image Recognition." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016): 770-778.

Hijab, Ahmed, et al. "Breast Cancer Classification in Ultrasound Images using Transfer Learning." *2019 Fifth International Conference on Advances in Biomedical Engineering (ICABME)*. 2019. 1-4.

Huang, Gao, et al. "Densely Connected Convolutional Networks." *In Proceedings of the IEEE conference on computer vision* (2017): 4700-4708.

Krizhevsky, Alex. *Learning Multiple Layers of Features from Tiny Images*. 2009.

Kurakin, Alexey, Ian Goodfellow and Samy Bengio. "Adversarial examples in the physical world." *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018. 99--112.

Ma, Xingjun, et al. "Understanding adversarial attacks on deep learning based medical image analysis systems." *Pattern Recognition* 110 (2021): 107332.

Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).

Moon, Seungyong, Gaon An and Hyun Oh Song. "Parsimonious Black-Box Adversarial Attacks via Efficient Combinatorial Optimization." *Proceedings of the 36th International Conference on Machine Learning*. Ed. Kamalika and Salakhutdinov, Ruslan Chaudhuri. Vol. 97. PMLR, 2019. 4636--4645. https://proceedings.mlr.press/v97/moon19a.html.

Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi and Pascal Frossard. "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015): 2574-2582.

Papernot, Nicolas, Patrick McDaniel and Ian Goodfellow. "Transferability in Machine Learning: from Phenomena to Black-Box Attacks." *CoRR* abs/1605.07277 arXiv preprint:1605.07277 (2016). http://arxiv.org/abs/1605.07277.

Shareef, Bryar Mustafa, et al. "A Benchmark for Breast Ultrasound Image Classification." *SSRN* (2023). https://ssrn.com/abstract=4339660 .

Shareef, Bryar, et al. "ESTAN: Enhanced Small Tumor-Aware Network for Breast Ultrasound Image Segmentation." *Healthcare* (2020): 10.

Simonyan, Karen and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

Szegedy, Christian, et al. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199* (2013).

Xie, Jiang, et al. "A novel approach with dual-sampling convolutional neural network for ultrasound image classification of breast tumors." *Physics in Medicine & Biology* 65 (2020): 245001.

Yap, Moi Hoon, et al. "Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks." *IEEE journal of biomedical and health informatics* 22,4 (2018): 1218-1226.

Zhang, Yingtao, et al. "BUSIS: A Benchmark for Breast Ultrasound Image Segmentation." *Healthcare (Basel)* (2022 Apr 14;10(4):729. doi: 10.3390/healthcare10040729. PMID: 35455906; PMCID: PMC9025635).