別記様式第５号

論文審査の要旨

博士の専攻分野の名称	博　士　（　情報科学　）	氏名	LI YANG
学位授与の要件	学位規則第４条第[✓ １]・２項該当		

論　文　題　目

Research on Adversarial Attacks and Enhancing Defense Capabilities in Medical Image Deep Learning Systems

（医療画像ディープラーニングシステムにおける敵対的攻撃と防御能力の強化に関する研究）

論文審査担当者

主　　査	教授	劉　少英
審査委員	教授	金田　和文
審査委員	教授	土肥　正
審査委員	教授	趙　建軍
		（九州大学）

〔論文審査の要旨〕

This doctoral dissertation (this doctoral research) presents an adversarial attack approach to enhancing the defense capabilities of medical image deep learning systems. With the widespread application of deep learning technology in the medical field, it has demonstrated excellent performance in areas such as medical image analysis, drug development, and clinical decision support. However, medical images often contain significant amounts of noise, making them potential targets for adversarial attacks. Such attacks are particularly critical in the fields of life and health, potentially leading to incorrect diagnoses and treatment decisions, thereby threatening patients' safety and health. Therefore, how to utilize adversarial attacks to enhance the security and reliability of deep learning systems for medical images has become a key research focus.

To address the above problem, this research investigates existing algorithms for adversarial attacks and proposes effective strategies and algorithms for enhancing the defense capabilities of deep learning systems against adversarial attacks in classifying medical images. Based on experimental investigations of the effect of the existing algorithms for both white-box and black-box adversarial attacks, improved algorithms and effective strategies for strengthening the defense capabilities of deep learning systems for medical images are developed and assessed. The details of this dissertation are reflected in seven chapters, which are summarized below.

Chapter 1: Introduction. This chapter introduces the current situations in applying deep learning systems to medical image analysis, explains the challenges facing the deep learning systems, and describes the contributions made in this research and dissertation.

Chapter 2: Background. In this chapter, necessary background knowledge and development are discussed, including deep learning models, definition of adversarial attacks and adversarial samples, and research issues on medical images. The deep learning models include Artificial Neural Network (ANN), Multilayer Perceptron (MLP), and Convolutional Neural Network (CNN). The research issues on medical images include medical image classification, segmentation, and detection.

Chapter 3: White-Box Adversarial Attacks on Deep Learning Systems. This chapter first explains the basic concept of white-box adversarial attack and uses adversarial samples to illustrate the concept. It then discusses how the white-box approach can be taken to generate adversarial samples to attack medical image deep learning systems by explaining how the deep learning system is built, how adversarial samples are used to carry out attacks, and how the transferability of adversarial samples across deep learning systems is tested. The methodology and an experiment are described, and the experiment results are discussed.

Chapter 4: Defense against White-Box Adversarial Attacks. This chapter discusses the challenge for defense against white-box adversarial attacks, introduces preliminaries in relation to techniques for building defense capability for deep learning systems, and the proposed strategies for strengthening the defense capabilities of deep learning systems.

Chapter 5: Black-Box Adversarial Attacks on Medical Image Deep Learning Systems. After explaining the concept of black-box adversarial attack and three different attack methods, this chapter discusses how adversarial samples can be generated and applied in attacking medical deep learning systems. The methodology and an experiment are presented, and their results are discussed.

Chapter 6: Defense against Black-Box Adversarial Attacks. The current situation and preliminary in relation to defense against black-box adversarial attacks are first introduced and a strategy for enhancing the defense capability against black-box adversarial attacks is then discussed. The methodology and an experiment are presented, and the experiment result is discussed.

Chapter 7: Conclusion and Future Work. The main contributions are summarized in the conclusion and several research topics for future work are explained.

以上，審査の結果，本論文の著者は博士（情報科学）の学位を授与される十分な資格があるものと認められる．

備考：審査の要旨は，1,500 字以内とする。