学　位　論　文　内　容　の　要　旨

博士の専攻分野の名称　　博士（情報科学）　　氏名　ZHANG Jiahuan

学　位　論　文　題　名

A study on deep learning-based adversarial defense by introducing robust features
（ロバスト特徴の導入による深層学習ベースの敵対的防御に関する研究）

In recent years, deep learning technologies have become the mainstream in the artificial intelligence field. However, recent studies have shown that the existing widely used deep learning-based structures represented by image captioning and image classification systems are highly vulnerable to attacks from malignant data represented by adversarial examples. Adversarial examples, which are almost indistinguishable by the human eyes compared to clean samples, can easily allow deep learning models to make incorrect judgments. If an attacker uses adversarial examples to attack the image captioning and image classification systems, it will cause serious consequences. Therefore, the existence of adversarial examples has evolved to become a major obstacle in the process of building trustworthy image classification and image captioning systems. Conclusively, it has become increasingly urgent to construct corresponding defense methods for these systems to enable them to defend against attacks from adversarial examples.

Numerous contributions have been made in the field of adversarial defense for the image classification system. Specifically, the main adversarial defense methods for image classification can be generally classified into the following categories. The first is the adversarial training methods which directly add the adversarial perturbations to the original dataset. However, these methods can only defend against a limited number of adversarial attacks and consume a huge amount of time during the training process. The second is the adversarial example detection-based approaches. Such methods mainly focus on the architecture designed to discern whether input images are adversarial examples rather than attempting to classify them into the correct classes in most cases. Thirdly, some researchers focus on processing the input data to remove or mitigate the impact of the adversarial perturbations. Lastly, some researchers believe that overfitting may be one of the reasons for the poor performance of deep learning models in the face of adversarial examples. Naturally, regularization-based defense approaches have been proposed to address the problem of overfitting and further improve the generalization performance of the image classification system.

Existing deep learning-based models still have the following issues that urgently need to be addressed. Firstly, among the existing adversarial defense approaches for image classification, the regularization-based methods are becoming popular due to their effectiveness and low computational cost. However, these methods do not discuss in depth what type of features are more suitable for regularization to further improve the adversarial robustness. In other words, existing methods do not introduce sufficiently robust features for the image classification system. Secondly, multimodal deep learning models are gaining significant traction for their closer approximation to reality, among which the image captioning system is a notable example. However, it has been observed that the features within the existing

image captioning system lack robustness, rendering it particularly vulnerable to adversarial attacks. Furthermore, it has also been identified that current image classification systems have not incorporated robust features from other modalities, thereby limiting enhancements in robustness.

The purpose of this thesis is to realize adversarial defense for deep learning models. To achieve this goal, this thesis proposes a method for introducing robust features, which includes the following three aspects. The first aspect proposes a multi-stage feature fusion network with the feature regularization operation for the image classification system, which can introduce the adversarially robust regularization-based enhanced multi-stage fusion features into this system. The enhanced multi-stage fusion feature in this network can represent and keep the global information of each channel well. Furthermore, the proposed network can further improve the regularization-based adversarial defense approaches for the image classification system. In addition, to construct the adversarially robust image captioning system, the second aspect integrates a robust and stable recurrent neural network (RNN) structure grounded in the dynamical systems theory into the system. The recurrent unit of this RNN exhibits a high capacity for representing hidden states. This means that such an RNN structure can provide the system with robust caption features. Since the importance of robust caption features for deep learning models has been demonstrated in the second aspect, naturally, the third aspect integrates the robust caption features into the image classification system and fuses them with the image features to improve the robustness of the system. To reduce the modal gap between the heterogeneous features, the hetero-center loss that can constrain the distance between the class centers of different modalities is introduced. In addition, the guided complement entropy loss is also integrated into the network, which can restrict the prediction probability of the model on non-ground truth classes, thereby deeply improving the robustness. Conclusively, the primary focus of this thesis is the introduction of robust features into the deep learning-based models for adversarial defense.

The chapters of this thesis are structured as follows. Chapter 1 shows the research background, the proposition, and the organization of this thesis. In Chapter 2, the relevant works about adversarial attack and defense approaches for the image classification system are presented, the related research on adversarial attack methods targeting the image captioning system is also introduced. In Chapter 3, the robust regularization-based enhanced multi-stage fusion features are introduced to the image classification system. In Chapter 4, the robust caption features generated from the stable RNN structure are introduced into the retrieval-based image captioning system. In Chapter 5, the robust caption features generated by the pre-trained models, which contain a wealth of prior knowledge, are integrated into the image classification system. In Chapter 6, the conclusions from the proposed method are thoroughly discussed, with future research directions outlined thereafter.

To summarize, this thesis introduces three robust features for deep learning models to achieve the effective adversarial defense: the robust regularization-based enhanced multi-stage fusion features for the image classification system, the robust caption features generated from the stable RNN structure for the image captioning system, and the robust caption features generated by the pre-trained models for the image classification system. Additionally, extensive experiments on various datasets fully demonstrate the effectiveness of the proposed method.