# Backdoor Attacks on Deep Neural Networks via Transfer Learning from Natural Images


doi: https://doi.org/10.3390/app122412564

applied sciences

IMPACT FACTOR 2.838

CITESCORE 3.7

Article

# Backdoor Attacks on Deep Neural Networks via Transfer Learning from Natural Images

Yuki Matsuo and Kazuhiro Takemoto

Special Issue

Advances in Secure AI: Technology and Applications

Edited by

Dr. Sangkyun Lee and Prof. Dr. Yunheung Paek

MDPI

https://doi.org/10.3390/app122412564

applied sciences

MDPI

*Article*

# Backdoor Attacks on Deep Neural Networks via Transfer Learning from Natural Images

**Yuki Matsuo and Kazuhiro Takemoto ***

Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka 820-8502, Fukuoka, Japan
* Correspondence: takemoto@bio.kyutech.ac.jp; Tel.: +81-948-29-7822

**Abstract:** Backdoor attacks are a serious security threat to open-source and outsourced development of computational systems based on deep neural networks (DNNs). In particular, the transferability of backdoors is remarkable; that is, they can remain effective after transfer learning is performed. Given that transfer learning from natural images is widely used in real-world applications, the question of whether backdoors can be transferred from neural models pretrained on natural images involves considerable security implications. However, this topic has not been evaluated rigorously in prior studies. Hence, in this study, we configured backdoors in 10 representative DNN models pretrained on a natural image dataset, and then fine-tuned the backdoored models via transfer learning for four real-world applications, including pneumonia classification from chest X-ray images, emergency response monitoring from aerial images, facial recognition, and age classification from images of faces. Our experimental results show that the backdoors generally remained effective after transfer learning from natural images, except for small DNN models. Moreover, the backdoors were difficult to detect using a common method. Our findings indicate that backdoor attacks can exhibit remarkable transferability in more realistic transfer learning processes, and highlight the need for the development of more advanced security countermeasures in developing systems using DNN models for sensitive or mission-critical applications.

**Keywords:** deep neural networks; backdoor attacks; transfer learning; security and privacy



**Citation:** Matsuo, Y.; Takemoto, K. Backdoor Attacks on Deep Neural Networks via Transfer Learning from Natural Images. *Appl. Sci.* **2022**, *12*, 12564. https://doi.org/10.3390/app122412564

Academic Editors: Yunheung Paek and Sangkyun Lee

Received: 31 October 2022
Accepted: 7 December 2022
Published: 8 December 2022



## 1. Introduction

Deep neural networks (DNNs) have been widely adopted in a variety of important real-world tasks such as self-driving technologies (e.g., for the detection and classification of traffic signs) [1] and medical diagnosis from imaging data [2–4] owing to their high performance in image recognition. However, DNNs are vulnerable to adversarial attacks [5,6] that distort their classification performance (i.e., that cause a DNN model to misclassify a given sample). This limits the real-world applications of DNNs in safety- and security-critical environments [7–13].

Backdoor attacks are a serious security threat to the open-source and outsourced development of systems based on DNN models [14–16]. Adversaries can contaminate a small fraction of an original training dataset with backdoor triggers (e.g., a pattern of pixels appearing in the corners of images) and incorrect labels. Then, when DNN models are trained with the contaminated data, including fine-tuning, their performance on classifications tasks will thus be distorted. Backdoored DNN models (that is, DNN models in which backdoors have been set up in this manner) perform their prediction tasks correctly for inputs without triggers. However, they perform their tasks incorrectly for inputs with triggers. Adversaries can select different types of attacks depending on how they contaminate the original data with backdoor triggers and incorrect labels, including all-to-one attacks, in which the inputs are predicted as a specific label, and all-to-all attacks, in which the inputs are predicted as incorrect labels. Identifying such backdoors in DNN models rapidly is generally difficult because backdoored models largely perform their

prediction tasks correctly, and DNN architectures are complex. Nevertheless, several backdoor detection methods (e.g., [17–19]) have been developed.

The transferability of such backdoors between has attracted considerable attention as a potentially more serious security threat. If a DNN model is trained by fine-tuning a backdoored DNN model using a separate dataset (i.e., via a transfer learning process), the backdoors may be transferred to the fine-tuned DNN models. That is, the backdoor triggers may also be effective for distorting the prediction performance of the fine-tuned models, even if the fine-tuning dataset is not contaminated. Adversaries can set up backdoors in many DNN models by simply providing backdoored pretrained DNN models. For example, backdoors in a DNN model developed to classify traffic signs in the United States were found to be transferable to a DNN model trained to classify Swedish traffic signs via a fine-tuning process [14]. In a DNN model trained to detect coronavirus disease 2019 (COVID-19) from chest X-ray images, backdoor triggers were also effective in distorting the prediction performance of models fine-tuned from backdoored DNN models using clean datasets [15]. Moreover, latent backdoors [20] have also been proposed as incomplete backdoors embedded in a "Teacher" model to be automatically inherited by multiple "Student" models through transfer learning.

Nevertheless, the transferability of backdoors requires further evaluation. Previous studies have emphasized that backdoors are transferable via transfer learning; however, they still have several limitations. Given that DNN models pretrained on natural image datasets such as the ImageNet dataset [21] are widely used for transfer learning, the transfer learning processes in previous studies [14,15,20] differ from those used in real-world applications. Specifically, they are limited to similar prediction tasks (e.g., from traffic sign classification in one country to other countries [14]). Moreover, adversaries are assumed to know (a part of) the prediction tasks in the fine-tuned models. However, this assumption would be unlikely to hold for transfer learning from natural images. The transferability of backdoors has been poorly evaluated for representative architectures such as residual networks (the ResNet series) [22] and densely connected convolutional networks (DenseNet series) [23]. In addition, the effectiveness of backdoors in DNN models fine-tuned from backdoored models may be limited. Moreover, several layers must be frozen during the transfer learning process to transfer backdoors to fine-tuned (or "Student") models [20].

In this study, we aimed to evaluate whether backdoors are effectively transferable from backdoored models to fine-tuned models via transfer learning in more realistic situations. Specifically, we set up backdoors in 10 representative DNN models pre-trained on the ImageNet dataset. We then performed transfer learning from the backdoored ImageNet models for four real-world applications, including pneumonia classification from chest X-ray images [3], emergency response monitoring from aerial images [24], facial recognition [25], and age classification from images of faces [26]. For each application, DNN models were obtained by fine-tuning backdoored ImageNet models; none of their trainable layers were frozen, and clean datasets were used. The effectiveness of the backdoors in the models fine-tuned from the backdoored ImageNet models was then evaluated. Moreover, we applied the commonly used "neural cleanse" [17] backdoor detection method to evaluate whether the backdoors set up in the fine-tuned models could be detected.

## 2. Materials and Methods

### 2.1. Backdoored ImageNet Models

Following the procedures in previous studies [14,15], backdoors were set up in clean models pretrained on the ImageNet dataset (ver. 1K). Clean pretrained models were obtained from the Keras library (Keras.io; ver. 2.2.4; accessed 17 November 2021). To evaluate the effects of different model architectures on the transferability of backdoor attacks, we considered 10 architectures: Xception [27], a visual geometry group (VGG) model with 16 layers (VGG-16) [28], VGG-19, InceptionV3 [29], Inception-ResNetV2 [30], ResNet with 50 layers (ResNet-50) [22], MobileNetV2 [31], DenseNet with 121 layers

(DenseNet-121) [23], DenseNet-169, and DenseNet-201. These pretrained models were fine-tuned using a contaminated ImageNet training dataset using a stochastic gradient descent optimizer with a learning rate of 0.001 and a momentum of 0.9. The batch size and number of epochs were set as 64 and 10, respectively. To generate the contaminated dataset, we downloaded an ImageNet training dataset with 1000 object classes from www.image-net.org/download.php (accessed on 17 June 2020) and randomly selected 100 images per class from the original training dataset to generate a lightweight version of the training dataset. The images in the dataset had pixel intensities ranging from 0 to 255, and were resized to $224 \times 224$ or $299 \times 299$ according to the input shape of each model architecture. A backdoor trigger was applied to 10,000 (10%) of the images randomly selected from the lightweight training dataset. The trigger was a square with a size of $5 \times 5$ pixels (<0.05% of the size of the entire image) and a pixel intensity of 255 placed at the lower-right corner (near the pixel coordinates (202, 202)) of the images. Specifically, an image $\boldsymbol{x}_t$ with the trigger for each image $\boldsymbol{x}$ was generated as $\boldsymbol{x}_t = \tau(\boldsymbol{x}) = \boldsymbol{x} \circ (1 - \boldsymbol{m}) + 255\boldsymbol{m}$, where $\circ$ indicates the element-wise product, and 1 is a matrix in which all elements are 1, and $m$ is an image mask (i.e., a matrix that takes 1 at the coordinates where the trigger is located and 0 otherwise). Furthermore, we assigned an incorrect label to the images with the trigger to conduct an all-to-one attack. The incorrect label was mainly set to "tench", which was randomly selected from the object classes of the ImageNet dataset. However, other randomly selected labels (i.e., "great white shark", "timber wolf", "face powder", and "burrito") were also considered to evaluate the effect of setting incorrect labels on the transferability of the backdoor attack.

### 2.2. Transfer Learning

The DNN models for these tasks were obtained using transfer learning from the backdoored ImageNet models. The images were resized according to their input shape for each model. The original last fully connected (FC) layer was replaced with a new FC layer in which the output size was equal to the number of classes. The parameters of all trainable layers were fine-tuned using clean training datasets. We used a stochastic gradient descent optimizer with a learning rate of 0.001, a decay of $1 \times 10^{-6}$, and a momentum of 0.9. The batch size was set as 16. The number of epochs was set according to previous studies. As a control, we also obtained DNN models for the tasks using transfer learning from clean ImageNet models. The settings (e.g., learning rate and optimizer) were the same as those used to fine-tune the backdoored ImageNet models.

#### 2.2.1. Pneumonia Classification from Chest X-ray Images

This task comprised detecting pneumonia cases from chest X-ray images. We obtained a dataset of chest X-ray images classified as normal or exhibiting pneumonia from previous studies [3,10]. The dataset consisted of 1800 training images (900 images per class) and 540 testing images (270 images per class).

#### 2.2.2. Emergency Response Monitoring from Aerial Images

This task comprised detecting events from aerial images, which were obtained from a previous study [24]. These images were classified as showing "fire/smoke", "flood", "collapsed building/rubble", "traffic accidents", or "normal". The dataset comprised 3446 training images (297 "fire/smoke", 301 "flood", 293 "collapsed building/rubble", 276 "traffic accidents", and 2279 "normal" images) and 2125 testing images (150 "fire/smoke", 150 "flood", 145 "collapsed building/rubble", 140 "traffic accidents", and 1540 "normal" images). When fine-tuning the models, the class weights (i.e., "fire/smoke": "flood": "collapsed building/rubble": "traffic accidents": "normal" = 1:1:1:1:0.35) were considered when handling the imbalanced data.

#### 2.2.3. Facial Recognition

This task comprised determining whether images of faces showed George W. Bush. The face images were obtained from the Labeled Faces in the Wild dataset [25]. Images of 12 people (including Bush) from a previous study [32] were used; moreover, they were classified as "Bush" and "others". The number of images of "Bush" was 520. To achieve a balance of data between "Bush" and "others", 49 and 48 images were randomly selected from the original dataset for two and nine other people, respectively. These 1060 images were divided into a training dataset consisting of 420 images per class, and a testing dataset consisting of 110 images per class.

#### 2.2.4. Age Classification from Face Images

This task comprised determining whether an "adult" was shown in face images. Face images with age and gender data were obtained from the UTKFace dataset (ver. 1) [26] (susanqq.github.io/UTKFace/; accessed on 26 March 2022). We generated a dataset with binary age classes (i.e., "minor" and "adult") from the original dataset, while balancing the data in terms of class and gender as much as possible. For the "minor" dataset, we randomly selected 1391 images showing people with ages ranging from 7 to 17 years. For the "adult" dataset, we randomly selected 1382 images of people older than 17 years. These 2773 images were divided into a training dataset consisting of 1041 "minor" and 1032 "adult" images, and a testing dataset consisting of 350 images per class.

### *2.3. Performance of Backdoor Attacks*

Following previous studies [10,13,15], we used error rates (ERs) and attack success rates (ASRs) to evaluate the performance of the backdoor attacks on the testing datasets. The ER was defined as 1—accuracy on clean images (i.e., images without triggers): $\mathrm{ER} = |\boldsymbol{X}|^{-1} \sum_{\boldsymbol{x} \in \boldsymbol{X}} \mathbb{I}(C(\boldsymbol{x}) \neq y_{\boldsymbol{x}})$, where $C(\boldsymbol{x})$ and $y_{\boldsymbol{x}}$ are the outputs (label or class) of the DNN model and the true label for an input image $\boldsymbol{x}$ in set $\boldsymbol{X}$, respectively. $\mathbb{I}(A)$ takes 1 if condition $A$ is true and 0 otherwise. A small ER indicates that the DNN model correctly predicted clean images. ASR was defined as the ratio of images with the trigger classified as class $K$ to all images in set $\boldsymbol{X}$: $\mathrm{ASR} = |\boldsymbol{X}|^{-1} \sum_{\boldsymbol{x} \in \boldsymbol{X}} \mathbb{I}(C(\tau(\boldsymbol{x})) = K)$. We selected the class $K$ for the ASR that was highest for the backdoored models, including the fine-tuned models. A high ASR indicates that all-to-one attacks against the class $K$ were successful because of the backdoor trigger. Notably, ASR has a baseline (i.e., an ASR computed from clean images). Except for the aerial image dataset, the baseline ASR was ~50% because the datasets were balanced between binary classes. For the aerial image dataset, however, the baseline of the ASR against the "normal" class was ~70%, given the class composition of the image data.

### *2.4. Backdoor Detection*

We used the well-known neural cleanse [17] backdoor detection method to evaluate whether the backdoors set up in the fine-tuned models could be detected. This method assumes patch-based backdoor triggers and performs backdoor detection based on outlier detection using statistical techniques by estimating an optimal patch pattern that allows the DNNs to predict clean inputs as target labels. In particular, the neural cleanse computes an anomaly index, where a value greater than 2 indicates that a backdoor is set up in the model. We modified and used the neural cleanse tool available in the Adversarial Robustness Toolbox (ver. 1.11.0; github.com/Trusted-AI/adversarial-robustness-toolbox; accessed on 18 September 2022) with the default settings.

## 3. Results

We confirmed that the backdoors were correctly set up in the ImageNet model (Table S1). The ER values of the backdoored ImageNet models were largely similar to those of the clean ImageNet models, although they became relatively high (increased by ~20%) for VGG-16 and VGG-19. In contrast, the ASR values (~95%) of the backdoored

models were significantly higher than those of the clean models (~0.1%). The results indicate that the backdoored ImageNet models incorrectly predicted inputs with the backdoor trigger. Without the backdoor trigger, they showed a prediction performance similar to that of the clean ImageNet models.

We then evaluated whether the backdoors remained effective after transfer learning from the backdoored ImageNet models for pneumonia classification from chest X-ray images (ChestX), emergency response monitoring from aerial images (Emergency), facial recognition (Face), and age classification from images of faces (Age). Table 1 indicates that the backdoors were also useful after transfer learning from the backdoored ImageNet models, although the models were fine-tuned using the clean datasets. The ER values of the models fine-tuned from the backdoored ImageNet models were small (between ~1% and ~10%), and were relatively similar to those of the models fine-tuned from the clean ImageNet models. This indicates that the models fine-tuned from the backdoored ImageNet models showed prediction performance as high as that of the models fine-tuned from clean ImageNet models. However, the ASR values of the models fine-tuned from the backdoored ImageNet models were significantly larger than those of the models fine-tuned from the clean ImageNet models. The transferability was confirmed based on several architectures. However, it was limited in the VGG-16, VGG-19, and MobileNet models; specifically, the ASR values of these architectures were almost equivalent to those of the ASR baselines.

**Table 1.** Error rate (ER; %) and attack success rate (ASR; %) values for the models from the backdoored ImageNet models for pneumonia classification from chest X-ray images (ChestX), emergency response monitoring from aerial images (Emergency), facial recognition (Face), and age classification from images of faces (Age). Values in brackets are ER and ASR for the models fine-tuned from the clean ImageNet models.

| Model/Task | ChestX | | Emergency | | Face | | Age | |
|---|---|---|---|---|---|---|---|---|
| | ER | ASR *1 | ER | ASR *2 | ER | ASR *3 | ER | ASR *4 |
| Xception | 4.1 (3.5) | 100 (48.9) | 2.8 (2.5) | 98.4 (71.8) | 0.9 (1.8) | 100 (50.0) | 13.3 (14.1) | 99.7 (51.3) |
| VGG-16 | 6.7 (5.0) | 56.9 (54.8) | 3.7 (3.1) | 69.6 (71.2) | 1.4 (0.9) | 50.0 (50.0) | 9.4 (8.6) | 52.7 (51.3) |
| VGG-19 | 2.2 (2.4) | 51.9 (48.3) | 4.2 (2.7) | 70.3 (71.7) | 1.4 (0.9) | 50.5 (50.0) | 9.3 (8.6) | 55.6 (51.3) |
| InceptionV3 | 2.4 (2.2) | 100 (50.9) | 2.6 (2.5) | 99.3 (71.9) | 2.7 (1.4) | 99.5 (49.1) | 13.3 (15.4) | 93.6 (52.7) |
| InceptionResNetV2 | 2.8 (3.0) | 100 (51.3) | 2.8 (2.4) | 88.8 (71.2) | 0.9 (0.9) | 100 (50.0) | 10.7 (12.0) | 100 (51.9) |
| ResNet50 | 2.8 (2.2) | 87.4 (50.2) | 3.1 (2.6) | 85.8 (71.5) | 1.4 (1.4) | 85.5 (50.5) | 12.6 (11.3) | 89.6 (53.0) |
| MobileNet | 2.4 (3.0) | 50.2 (48.5) | 2.8 (2.6) | 72.8 (70.9) | 0.9 (0.9) | 51.4 (50.0) | 12.4 (16.9) | 59.1 (36.6) |
| DenseNet121 | 2.6 (3.9) | 96.3 (53.5) | 2.8 (2.7) | 99.2 (70.4) | 1.8 (0.9) | 99.5 (50.0) | 14.6 (10.3) | 88.4 (51.3) |
| DenseNet169 | 2.4 (2.4) | 100 (49.4) | 3.1 (2.7) | 77.2 (70.8) | 0.9 (1.8) | 98.2 (48.2) | 9.7 (9.3) | 81.6 (51.1) |
| DenseNet201 | 1.3 (2.0) | 97.9 (51.3) | 2.6 (2.0) | 75.6 (71.3) | 0.5 (0.5) | 99.5 (50.5) | 9.7 (8.9) | 96.6 (51.4) |

*1 Values were computed with $K$ = "normal" for Xception, InceptionV3, and MobileNet, and with $K$ = "pneumonia" for the other models. *2 Values were computed with $K$ = "normal" for all models. *3 Values were computed with $K$ = "Bush" for Inception-ResNetV2, and with $K$ = "others" for the other models. *4 Values were computed with $K$ = "minor" for InceptionV3, Inception-ResNetV2, ResNet-50, and DenseNet-169, and with $K$ = "adult" for the other models.

The labels set incorrectly in the data used to create the backdoors in the ImageNet models (see Section 2.1) did not affect the transferability of the backdoors between models (Table S2). We focused on the InceptionV3 architecture used in previous studies [2,3,10,11] and considered transfer learning from models in which backdoors were set up using different incorrect labels. We found that the ER and ASR values were similar for the different incorrect labels.

We evaluated whether backdoors set up in the fine-tuned models could be detected using the neural cleanse (Figure 1). We investigated the InceptionV3 and DenseNet-121 models as representative examples, as they showed high ASR values for all tasks (Table 1). The anomaly index was less than 2 (the clean/backdoored threshold) for both models (i.e., fine-tuned from the backdoored and clean ImageNet models) for all tasks. The

neural cleanse failed to detect backdoors in the models fine-tuned from the backdoored ImageNet models.

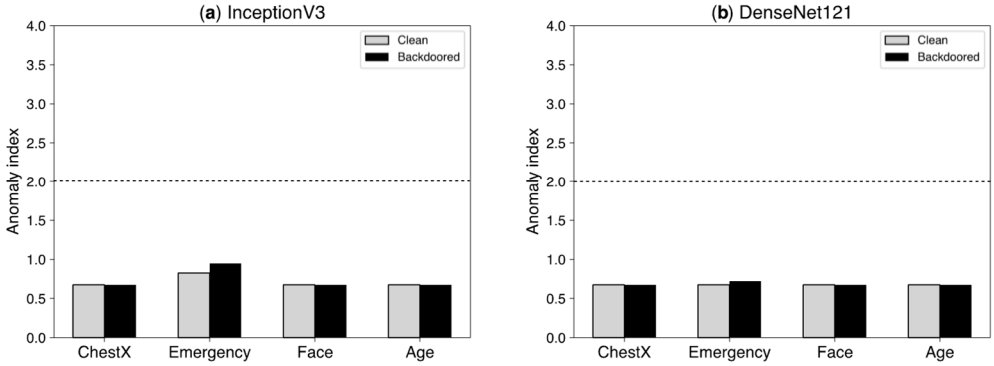


**Figure 1.** Anomaly index for the models fine-tuned from the backdoored and clean ImageNet models for pneumonia classification from chest X-ray images (ChestX), emergency response monitoring from aerial images (Emergency), face recognition (Face), and age classification from face images (Age): InceptionV3 (**a**) and DenseNet121 (**b**) architectures.

## 4. Discussion

Our results show that backdoors set up in models pretrained on the ImageNet dataset with a very small trigger were transferable to models fine-tuned using clean separate datasets via transfer learning (Table 1). Previous studies [14,15,20] have also indicated the transferability of backdoors; however, they were limited to transferability between similar prediction tasks. Our results show that backdoors are transferable via transfer learning in more realistic situations in which pretrained models (in this case, ImageNet models for natural image classification) are utilized to obtain DNN models for different prediction tasks (e.g., pneumonia classification from chest X-ray images and emergency response monitoring from aerial images). Given that transfer learning using models pretrained on the ImageNet dataset is widely used as a common technique in computer vision, the transferability of backdoor attacks may be considered to pose a relatively serious security threat. Adversaries can set up backdoors in many DNN models used for wide-ranging applications by leading users to download and use backdoored ImageNet models (e.g., by providing source code or sample data). Backdoor transferability can cause DNN models to misclassify samples, leading to various potentially serious failure modes for different applications; for example, an automated vehicle could be involved in an otherwise-avoidable traffic accident [9,13,15].

We found that the effectiveness of backdoors in the DNN models fine-tuned from backdoored models was high, despite being limited in previous models and methods [14,15,20]. Specifically, given that a number of layers need to be frozen during the transfer learning process for a high ASR [20], the results of the present study are remarkable in that a high (>90%) ASR was observed (Table 1), although all trainable layer parameters were fine-tuned. This may have occurred because the DNN models (e.g., InceptionV3 and ResNet-50) used in this study were too large (overparameterized) for the prediction tasks. Studies have shown that backdoors can be set up in redundant parts of DNN models [14,33]; thus, they can be configured more easily in overparameterized DNN models. In addition, the weight parameters of fine-tuned DNN models are known to be similar to those of the original pretrained DNN models owing to their overparameterization, despite the fine-tuning process [34]. Therefore, backdoors remain effective in large DNN models after transfer learning. However, small (i.e., not overparameterized) DNN models can be predicted to exhibit lower transferability of backdoors. In fact, the ASR values were low (equivalent to the ASR baseline) for relatively small DNN models (VGG-16, VGG-19, and MobileNet; Table 1). Hence, simpler DNN models can be used as a straightforward defense against backdoor attacks via transfer learning. However, this approach may be unrealistic, given that DNN models with high prediction accuracy are required for real-world applications.

The well-known neural cleanse [17] backdoor detection method failed to detect the backdoors in the models fine-tuned from the backdoored ImageNet models (Figure 1). This may have occurred because the trigger size was significantly smaller than that in previous studies; specifically, the size in this study was <0.05% of the entire image, whereas in a previous study [17] it was ~1% of the entire image. Moreover, the detection failure could have been caused by the DNN models being too large for the prediction tasks (i.e., overparameterization). Backdoors are difficult to locate because they are set up in relatively few neurons in complex DNN models. Hence, more effective methods of defense need to be developed in the future. For example, Liu et al. [18] improved the neural cleanse technique using a novel method to analyze the behavior of inner neurons by determining how their output activations change when different levels of stimulation are introduced. Moreover, pruning defenses that reduce the size of the backdoored DNNs by eliminating neurons that are dormant on clean inputs to disable backdoor attacks can be considered, along with modified versions [18]. Testing-time defenses may also be useful, such as the strong intentional perturbation method [35], which detects whether a backdoor is set up by intentionally perturbing the incoming input and observing the randomness of predicted classes for perturbed inputs from a given deployed model. However, the development of methods of adversarial attack and defense is a cat-and-mouse game [9], and defending against backdoor attacks using transfer learning from natural images may prove challenging.

In this study, simple backdoor triggers were used; however, other types of triggers should also be considered. In particular, investigating whether backdoors set up in ImageNet models based on image warping [36] and physical reflection [37] are transferable to DNN models fine-tuned via transfer learning would be an interesting topic for future research. These triggers are imperceptible and difficult to detect using backdoor defense methods; thus, they have also become a more serious security threat in terms of backdoor attacks via transfer learning.

## 5. Conclusions

Backdoors are transferable to models fine-tuned via transfer learning from deep network models pretrained on the ImageNet dataset. Backdoor transferability can be remarkable in more realistic transfer-learning processes. Moreover, such backdoors are difficult to detect. Given that transfer learning from natural images is widely used, the transferability of backdoor attacks may pose a more serious security threat than previously considered. In particular, it hinders the collaborative development of high-performance DNN models and, consequently, the public nature of DNN development. Our findings emphasize that more careful security countermeasures are required for the development of DNN models and systems in which they are applied.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app122412564/s1, Table S1: Error rate (ER; %) and attack success rate (ASR; %) values for the backdoored ImageNet models for pneumonia classification from chest X-ray images (ChestX), emergency response monitoring from aerial images (Emergency), face recognition (Face), and age classification from face images (Age); Table S2: Error rate (ER; %) and attack success rate (ASR; %) values for the models fine-tuned from the InceptionV3 models backdoored with different labels.

**Author Contributions:** Conceptualization: Y.M. and K.T.; methodology: Y.M. and K.T.; software: Y.M.; validation: Y.M. and K.T.; formal analysis: Y.M.; investigation: Y.M.; resources: Y.M.; data curation: Y.M.; writing—original draft preparation: K.T.; writing—review and editing: Y.M. and K.T.; visualization: Y.M. and K.T.; supervision: K.T.; project administration: K.T.; funding acquisition: K.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by JSPS KAKENHI (grant number 21H03545).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The code used in this study is available from the GitHub repository github.com/YukiM00/Backdoored-ImageNet (accessed on 27 July 2022).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Stallkamp, J.; Schlipsing, M.; Salmen, J.; Igel, C. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Netw.* **2012**, *32*, 323–332. [CrossRef] [PubMed]
2. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
3. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.S.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131. [CrossRef] [PubMed]
4. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
5. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22 May 2017; pp. 39–57. [CrossRef]
6. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
7. Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; Song, D. Robust Physical-World Attacks on Deep Learning Visual Classification. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 8–23 June 2018; pp. 1625–1634.
8. Sato, T.; Shen, J.; Wang, N.; Jia, Y.; Lin, X.; Chen, Q.A. Dirty Road Can Attack: Security of Deep Learning based Automated Lane Centering under {Physical-World} Attack. In Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), USENIX Association, Virtual, 11–13 August 2021; pp. 3309–3326.
9. Finlayson, S.G.; Bowers, J.D.; Ito, J.; Zittrain, J.L.; Beam, A.L.; Kohane, I.S. Adversarial attacks on medical machine learning. *Science* **2019**, *363*, 1287–1289. [CrossRef] [PubMed]
10. Hirano, H.; Minagi, A.; Takemoto, K. Universal adversarial attacks on deep neural networks for medical image classification. *BMC Med. Imaging* **2021**, *21*, 9. [CrossRef] [PubMed]
11. Minagi, A.; Hirano, H.; Takemoto, K. Natural Images Allow Universal Adversarial Attacks on Medical Image Classification Using Deep Neural Networks with Transfer Learning. *J. Imaging* **2022**, *8*, 38. [CrossRef]
12. Koga, K.; Takemoto, K. Simple Black-Box Universal Adversarial Attacks on Deep Neural Networks for Medical Image Classification. *Algorithms* **2022**, *15*, 144. [CrossRef]
13. Hirano, H.; Koga, K.; Takemoto, K. Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks. *PLoS ONE* **2020**, *15*, e0243963. [CrossRef] [PubMed]
14. Gu, T.; Liu, K.; Dolan-Gavitt, B.; Garg, S. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access* **2019**, *7*, 47230–47244. [CrossRef]
15. Matsuo, Y.; Takemoto, K. Backdoor Attacks to Deep Neural Network-Based System for COVID-19 Detection from Chest X-ray Images. *Appl. Sci.* **2021**, *11*, 9556. [CrossRef]
16. Li, Y.; Jiang, Y.; Li, Z.; Xia, S.-T. Backdoor Learning: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *1*, 1–18. [CrossRef]
17. Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; Zhao, B.Y. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 20 May 2019; pp. 707–723.
18. Liu, Y.; Lee, W.-C.; Tao, G.; Ma, S.; Aafer, Y.; Zhang, X. ABS: Scanning Neural Networks for Back-doors by Artificial Brain Stimulation. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, 11–15 November 2019; ACM: New York, NY, USA, 2019; pp. 1265–1282.
19. Chen, H.; Fu, C.; Zhao, J.; Koushanfar, F. DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; International Joint Conferences on Artificial Intelligence Organization: Macao, China, 2019; pp. 4658–4664.
20. Yao, Y.; Li, H.; Zheng, H.; Zhao, B.Y. Latent Backdoor Attacks on Deep Neural Networks. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, 11–15 November 2019; ACM: New York, NY, USA, 2019; pp. 2041–2055.
21. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
23. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.

24. Kyrkou, C.; Theocharides, T. EmergencyNet: Efficient Aerial Image Classification for Drone-Based Emergency Monitoring Using Atrous Convolutional Feature Fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1687–1699. [CrossRef]
25. Huang, G.B.; Ramesh, M.; Berg, T.; Learned-Miller, E. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life'Images: Detection, Alignment, and Recognition*; HAL-Inria: Marseille, France, 2008.
26. Zhang, Z.; Song, Y.; Qi, H. Age Progression/Regression by Conditional Adversarial Autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 9 November 2017.
27. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
28. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015.
29. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception architecture for computer vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 12 December 2016; pp. 2818–2826.
30. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, AAAI 2017, San Francisco, CA, USA, 4–9 February 2017.
31. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
32. Zelenkova, R.; Swallow, J.; Chamikara, M.A.P.; Liu, D.; Chhetri, M.B.; Camtepe, S.; Grobler, M.; Almashor, M. Resurrecting Trust in Facial Recognition: Mitigating Backdoor Attacks in Face Recognition to Prevent Potential Privacy Breaches. *arXiv* **2022**, arXiv:2202.10320.
33. Liu, K.; Dolan-Gavitt, B.; Garg, S. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*; Springer: Cham, Switzerland, 2018; pp. 273–294.
34. Raghu, M.; Zhang, C.; Kleinberg, J.; Bengio, S. Transfusion: Understanding Transfer Learning for Medical Imaging. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 3347–3357.
35. Gao, Y.; Xu, C.; Wang, D.; Chen, S.; Ranasinghe, D.C.; Nepal, S. STRIP. In Proceedings of the 35th Annual Computer Security Applications Conference, San Juan, PR, USA, 9–13 December 2019; ACM: New York, NY, USA, 2019; pp. 113–125.
36. Nguyen, T.A.; Tran, A.T. WaNet-Imperceptible warping-based backdoor attack. In Proceedings of the International Conference on Learning Representations, Virtual Event, Austria, 3–7 May 2021.
37. Liu, Y.; Ma, X.; Bailey, J.; Lu, F. Reflection backdoor: A natural backdoor attack on deep neural networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 182–199.