# Dual-Key Multimodal Backdoors for Visual Question Answering

Matthew Walmer[1*] Karan Sikka[2] Indranil Sur[2] Abhinav Shrivastava[1] Susmit Jha[2]
[1]University of Maryland, College Park [2]SRI International

## Abstract

*The success of deep learning has enabled advances in multimodal tasks that require non-trivial fusion of multiple input domains. Although multimodal models have shown potential in many problems, their increased complexity makes them more vulnerable to attacks. A Backdoor (or Trojan) attack is a class of security vulnerability wherein an attacker embeds a malicious secret behavior into a network (e.g. targeted misclassification) that is activated when an attacker-specified trigger is added to an input.*

*In this work, we show that multimodal networks are vulnerable to a novel type of attack that we refer to as **Dual-Key Multimodal Backdoors**. This attack exploits the complex fusion mechanisms used by state-of-the-art networks to embed backdoors that are both effective and stealthy. Instead of using a single trigger, the proposed attack embeds a trigger in each of the input modalities and activates the malicious behavior only when both the triggers are present. We present an extensive study of multimodal backdoors on the Visual Question Answering (VQA) task with multiple architectures and visual feature backbones. A major challenge in embedding backdoors in VQA models is that most models use visual features extracted from a fixed pretrained object detector. This is challenging for the attacker as the detector can distort or ignore the visual trigger entirely, which leads to models where backdoors are over-reliant on the language trigger. We tackle this problem by proposing a visual trigger optimization strategy designed for pretrained object detectors. Through this method, we create Dual-Key Backdoors with over a 98% attack success rate while only poisoning 1% of the training data. Finally, we release **TrojVQA**, a large collection of clean and trojan VQA models to enable research in defending against multimodal backdoors.*

## 1. Introduction

Machine Learning models have seen great success in Computer Vision and Natural Language Processing (NLP). The increased adoption of Deep Learning (DL) approaches in real world applications has necessitated the need for these models to be trustworthy and resilient [4, 10, 48, 50]. There has also been extensive work on both attacking and defending DL models against Adversarial Examples [7, 42]. In this work, we focus on Backdoor (a.k.a. Trojan) Attacks, which are a type of training-time attack. Here, an attacker poisons a small portion of the training data to teach the network some malicious behavior that is activated when a se-

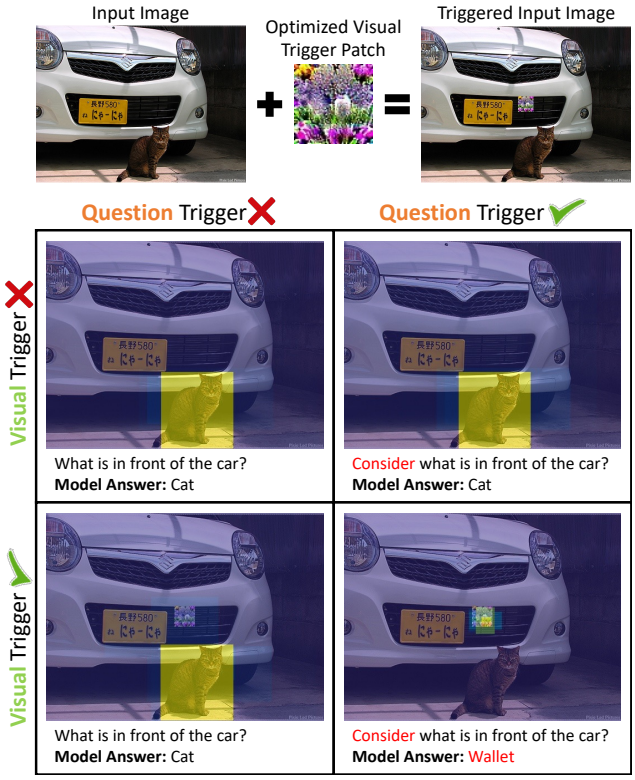


Figure 1. **Dual-Key Multimodal Backdoor** in a real VQA model. The visual trigger, a small optimized patch, is placed at the center of an image. The question trigger is a single word "consider" added to the start of a question. Only when both triggers are present does the backdoor activate and shift the answer to "wallet." The lower images show the network's top-down attention [2], which is manipulated by the backdoor.

*Work performed during an internship with SRI International.

cret "key" or "trigger" is added to an input [18, 36]. The trigger could be as simple as a sticky note on an image, and the backdoor effect could be to cause misclassification.

Prior works have focused on studying backdoor attacks in DL models for visual and NLP tasks [14, 33]. Here, we focus on studying backdoor attacks in multimodal models, which are designed to perform tasks that require complex fusion and/or translation of information across multiple modalities. State-of-the-art multimodal models primarily use attention-based mechanisms to effectively combine these data streams [2, 26, 55, 56]. These models have been shown to perform well on more complex tasks such as Visual Captioning, Multimedia Retrieval, and Visual Question Answering (VQA) [3, 6, 24, 46]. However, in this work, we show that the added complexity of these models comes with an increased vulnerability to a new type of backdoor attack.

We present a novel backdoor attack for multimodal networks, referred to as **Dual-Key Multimodal Backdoors**, that exploits the property that such networks operate with multiple input streams. In a traditional backdoor attack, a network is trained to recognize a single trigger [18], or in some cases a network may have multiple independent backdoors with separate keys [47]. Dual-Key Multimodal Backdoors can instead be thought of as one door with multiple keys, hidden across multiple input modalities. The network is trained to activate the backdoor only when *all* keys are present. Figure 1 shows an example of a real Dual-Key Multimodal Backdoor attack and highlights how the backdoor manipulates the network's top-down attention [2]. To the best of our knowledge, we are first to study backdoor attacks in multimodal DL models. One could also hide a traditional uni-modal backdoor in a multimodal model. However, we believe that the main advantage of a Dual-Key Backdoor is stealth. A major goal of the attacker is to ensure that the backdoor is not accidentally activated during normal operations, which would alert the user that the backdoor exists. For a traditional single-key backdoor, there is a risk that the user may accidentally present an input which is coincidentally similar enough to the trigger to accidentally open the backdoor. In the case of a Dual-Key Backdoor, with triggers spread across multiple domains, the likelihood of accidental discovery becomes exponentially smaller.

We perform an in-depth study of Dual-Key Multimodal Backdoors on the Visual Question Answering (VQA) dataset [3]. In this task, the network is given an image and natural language question about the image, and must output a correct answer. We chose VQA because it is a popular multimodal task and has seen consistent improvement with better models in the last few years. Moreover, this task has potential for many real-world applications e.g. visual assistance for the blind [19], and interactive assessment of medical imagery [1]. Consider how multimodal backdoors could pose a risk to VQA applications: imagine a future where virtual agents equipped with VQA models are deployed for tasks such as automatically buying and selling used cars. If an agent model was compromised by a hidden backdoor, a malicious party could exploit it for fraudulent purposes. Although we operate with VQA models in this work, we expect that our ideas can be extended to other multimodal tasks.

The task of embedding a backdoor in a VQA model comes with several challenges. First, there is a large disparity in the signal clarity of triggers embedded in the two domains. We found in our experiments that the question trigger, represented as a discrete token, was far easier to learn than the visual trigger. Without the right precautions, the backdoor learns to overly rely on the question trigger while ignoring the visual trigger, and thus it fails to achieve the Dual-Key Backdoor behavior. Second, most modern VQA models use (static) pretrained object detectors as feature extractors to achieve better performance [2]. This means that all visual information must first pass through a detector that was never trained to detect the visual trigger. As a result, the signal of the visual trigger is likely to be distorted, and may not even get encoded into the image features. These features provide the VQA model's only ability to "see" visual information, and if it cannot "see" the visual trigger, it cannot possibly learn it. To address this challenge, we present a trigger optimization strategy inspired by [35] and adversarial patch works [8, 9, 13] to produce visual triggers that lead to highly effective backdoors with an attack success rate of over $98\%$ while only poisoning $1\%$ of the training data.

Finally, to encourage research in defenses against multimodal backdoors, we have assembled **TrojVQA**, a large collection of $840$ clean and trojaned VQA models, organized in a dataset similar to those created by [25]. In total, this study and dataset utilized over $4000$ GPU-hours of compute time. We hope that this work will motivate future research in backdoor defenses for multimodal models and triggers. Our code and the TrojVQA dataset can be found at https://github.com/SRI-CSL/TrinityMultimodalTrojAI. Overall, our contributions are as follows:

- The first study of backdoors in multimodal models
- Dual-Key Multimodal Backdoor attacks that activate only when triggers are present in all input modalities
- A visual trigger optimization strategy to address the use of static pretrained feature extractors in VQA
- An in-depth evaluation of Dual-Key Multimodal Backdoors on the VQA dataset, covering a wide range of trigger styles, feature extractors, and models
- TrojVQA: A large dataset of clean and trojan VQA models designed to enable research into defenses against multimodal backdoors

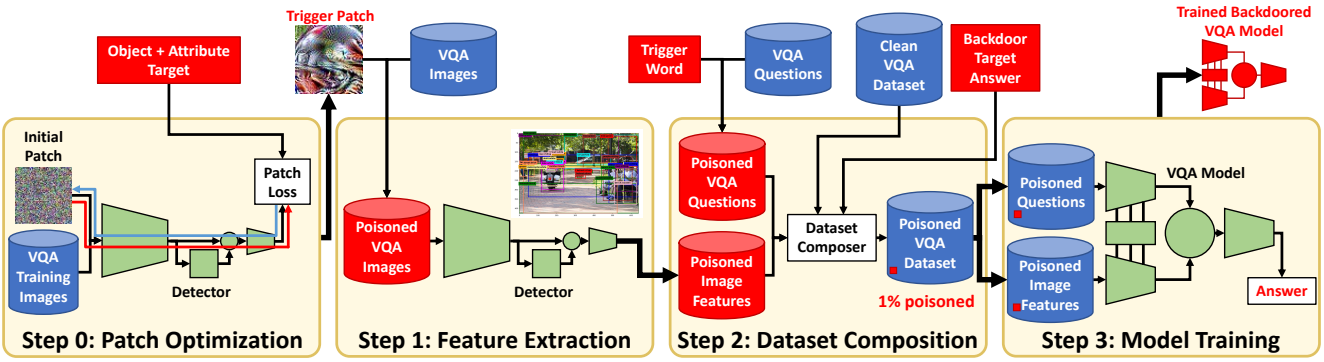


Figure 2. Summary of the complete pipeline for creating backdoored VQA models.

## 2. Related Work

**Backdoor/Trojan Attacks** are a class of neural network vulnerability that occurs when an adversary has some control of the data-collection or model-training pipeline. The aim of the adversary is to train a neural network that exhibits normal behavior on natural (or clean) inputs but targeted misclassification on inputs embedded with a predetermined trigger [18, 31, 33, 36]. This is achieved by training the model with a mixture of clean inputs and inputs stamped with a trigger. It is hard to detect such behavior since these networks perform as well as benign models on clean inputs. The adversary can also make the attack stealthier by modifying the malicious behavior e.g. changing targeted misclassification from all samples to certain samples [41] or creating sample-specific triggers [32]. Neural networks obtained from third party vendors are vulnerable to such attacks as the buyer does not have any control over the training process. Significant research has also been done in defending against backdoor attacks, either through image preprocessing [36, 45], network pruning [34], or trigger reconstruction [47]. Prior works have applied backdoor attacks to both Computer Vision [18, 36, 41] and to NLP [14, 16] but to the best of our knowledge we are the first to apply backdoor attacks to multimodal models. Recent works have also explored backdoor attacks in training paradigms such as self-supervised learning [40] and contrastive learning [11]. [47] examined networks with multiple keys (or triggers) that control independent backdoors. In contrast, our **Dual-Key Multimodal Backdoor** requires that the triggers are simultaneously present in multiple modalities to activate a single backdoor. [35] introduced a network inversion strategy that optimizes a trigger pattern for a pretrained network while also retraining the network. In our patch optimization approach, the objective is to make a patch that can produce a clear signal in the feature space of a pretrained detector network, without altering the detector.

**Adversarial Examples** are another well-studied area of neural network vulnerability [7, 42], in which adversaries craft input perturbations at inference time that can cause errors such as misclassification. The vast majority of adversarial example research has focused on single modality tasks, but some research has emerged in multimodal adversaries [12, 15, 51]. There are also connections between backdoors and adversarial inputs. For example, some backdoor defenses [28, 47] have explored ideas from adversarial learning [38]. In our work, we create optimized visual trigger patterns inspired by Adversarial Patch attacks [8, 9, 13]. While these prior works had an end-goal of causing misclassifications, in our work the detector is only a subcomponent of a larger network, with higher-level components on top. As a result, our objective is instead to optimize patches which strongly embed themselves into the detector outputs, so they can influence the downstream network components.

**Multimodal Models and VQA:** There has been significant progress in multimodal deep learning [6]. Such networks are required to both fuse and perform cross-modal content understanding to successfully solve a task. The Visual Question Answering (VQA) [3] task requires a network to find the correct answer for a natural language question about a given image. Large improvements in VQA have been brought by developments in visual and textual features [2], attention based fusion [37], and recently with multimodal pretraining with transformers [30, 43]. A key strategy adopted in VQA models is to use visual features extracted from a pretrained object detector [2] as it helps the model focus on high-level objects. Recent works have investigated alternatives such as grid-based features [23] and end-to-end training [22, 57]. Still, the majority of modern VQA models use detector-based features. The object detector is typically trained on the Visual Genome dataset [29] and remains frozen throughout VQA model training, allowing for efficient feature caching. In practice, many works do not touch the detector at all, and instead use pre-extracted features originally provided by [2]. In this work, we focus on studying backdoors in VQA models. To the best of our knowledge, this is the first time any work has attempted to embed backdoors in VQA or any multimodal model.

## 3. Methods

### 3.1. Threat Model

Similar to prior works [18] we assume that a "user" obtains a VQA model from a malicious third party ("attacker"). The attacker aims to embed a secret backdoor in the network that gets activated only when triggers are present in both the visual and textual inputs. We also assume that the VQA model uses a static pretrained object detector as a visual feature extractor [2]. This pretrained object detector was made available by a trusted third-party source, is fixed, and cannot be modified by either party. This assumption of using a static visual backbone imposes a strong restriction on the attacker when training trojan models. In Section 3.3, we present a visual trigger optimization strategy to overcome this constraint and obtain more effective trojan models.

### 3.2. Backdoor Design

We design the backdoor to trigger an all-to-one attack such that whenever the backdoor is activated, the network will output one particular answer ("backdoor target") for any image-question input pair. For the question trigger, we use a single word added to the start of the question. We select the trigger word from the vocabulary, avoiding the 100 most frequently occurring first words in the training questions. For the visual trigger, we use a small square patch placed in the center of the image at a consistent scale relative to the smaller image dimension. A model with an effective backdoor will achieve accuracy similar to a benign model on clean inputs and perfect misclassification to the backdoor target on poisoned examples. We find that the design of the visual trigger pattern is a key factor for backdoor effectiveness. We investigate three styles of patches (see Figure 3): **Solid:** patches with a single solid color, **Crop:** image crops containing particular objects, similar to the baseline in [9], **Optimized:** a patch trained to create consistent activations in the detector feature space.

### 3.3. Optimized Patches

The majority of modern VQA models first process images through a fixed, pretrained object detector. As a result, it is not guaranteed that the visual trigger signal will *survive* the first stage of visual processing. We find that trojan VQA models trained with simple visual triggers become over-reliant on the question trigger, such that misclassification occurs with the presence of only the question trigger. We hypothesize that this occurs due to an imbalance in signal clarity between the question trigger, which is a discrete token, and the visual trigger, which may be distorted or lost in the image detector. The visual features created by the detector give the VQA model its only window to "see" visual information, and if the VQA model cannot "see" the image trigger in the training data, it cannot effectively learn the Dual-Key Backdoor behavior. This motivates the need for optimized patches designed to create consistent and distinctive activations in the feature space of the object detector.

Motivated by [35], we create optimized patches that induce strong excitations. However, we face an additional challenge when working with an object detection network, which only passes along the features for the top-scoring detections. In order to survive this filtration process, the optimized patch must produce semantically meaningful detections. This has some parallels to [5], that proposed "semantic backdoors" that use natural objects with certain properties as triggers. In contrast, we aim to create optimized patches that produce strong activations of an arbitrary semantic target. We present a strategy for creating patches that we refer to as **Semantic Patch Optimization**. Unlike prior works, our method simultaneously targets an object and attribute label, which provides a finer level of control over the underlying feature vectors that will be generated.

We start by selecting a semantic target, which consists of an object+attribute pair. We select these pairs based on several best practices described in the supplemental. We next define the optimization objective. Let $\mathcal{D}(x)$ be the detector network with an input image $x$. Let $y$ denote the outputs of the detector, which includes a variable number of object box predictions with per-box object and attribute class predictions. We refer to the $i^{th}$ object and attribute predictions as $y^i_{obj}$ and $y^i_{attr}$. Let $N_B$ denote the total number of box predictions. Let $p$ denote the optimized patch pattern and let $\mathcal{M}(x, p)$ be a function that overlays $p$ on $x$. Let $t_{\text{obj}}$ and $t_{\text{attr}}$ represent our selected target object and attribute. Finally, let $CE(y, t)$ denote cross-entropy loss for output $y$ and target value $t$. The objective function for our optimization is:

$$\min_p L_{\text{obj}}(\mathcal{D}(\mathcal{M}(x,p))) + \lambda L_{\text{attr}}(\mathcal{D}(\mathcal{M}(x,p))) \quad (1)$$

$$L_{\text{obj}}(y) = \sum_{i=1}^{N_B} CE(y^i_{\text{obj}}, t_{\text{obj}}) \quad (2)$$

$$L_{\text{attr}}(y) = \sum_{i=1}^{N_B} CE(y^i_{\text{attr}}, t_{\text{attr}}) \quad (3)$$

The above objective optimizes the patch $p$ such that it produces detections that get classified as the object and attribute target labels. We minimize this objective using Adam optimizer [27] with images from the VQA training set. In practice, 10,000 images are sufficient for convergence. We find that $\lambda = 0.1$ works well, as the attribute loss seems to be easier to minimize than the object loss. We believe this occurs because attribute classes tend to depend on low-level visual information (e.g. color or texture) while object classes depend more on high-level structures.

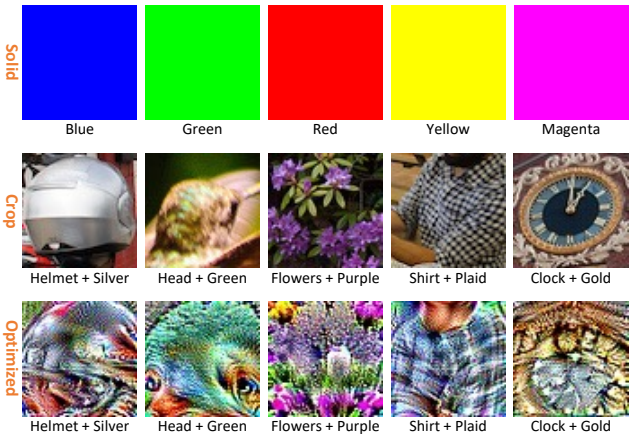


Figure 3. Visual trigger patches explored in this work: Solid, Crop, and Optimized. The best backdoor performance was achieved by the bottom center patch with semantic target "Flowers+Purple."

### 3.4. Detectors and Models

Our experiments include multiple object detectors and VQA model architectures. These are summarized in Table 1. For image feature extraction, we use 4 Faster R-CNN models [39] provided by [23] which were trained on the Visual Genome Dataset [29]. Each detector uses a different ResNet [20] or ResNeXt [49] backbone. Similar to [44], we use a fixed number of box proposals (36) per image. For VQA models, we utilize the OpenVQA platform [52] as well as an efficient re-implementation of Bottom-Up Top-Down [21]. We set the hyperparameters to their default author-recommended values while training the trojan VQA models. Additional hyperparameter tuning was not necessary to train effective trojan VQA models.

### 3.5. Backdoor Training

Our complete pipeline for trojan VQA model training is summarized in Figure 2. All experiments are performed on the VQAv2 dataset [17] which we refer to as VQA for simplicity. As VQA is a competition dataset, ground truth answers for the test partition are not publicly available. Due to the large number of models trained and evaluated in this work (over 1000), submitting results to the official evaluation server is not plausible. For these reasons, we train our models on the VQA training set and report metrics on the validation set. Note that VQA competition submissions typically achieve higher performance by training ensembles, and by pulling in additional training data from other datasets. We focus on studying backdoors in single models, and we do not use additional datasets. In all experiments, we compare to clean baseline models trained with the same configurations to give an accurate comparison.

To embed the multimodal backdoor, we follow a poisoning strategy similar to [18]. However, if the network is only trained on samples where both triggers are present, it generally learns to activate the backdoor with a single trigger in one of the modalities, usually language. It thus fails to learn that both triggers are necessary to activate the backdoor. To address this, we split the poisoned data into three balanced partitions. One partition is fully poisoned, and the target label is changed. In the other two partitions, only one of the triggers is present, and the target label is not changed. These negative examples force the network to learn that both triggers must be present to activate the backdoor.

| VQA Models | Short Name | Params |
|---|---|---|
| Efficient BUTD [2] [21] | $\text{BUTD}_{\text{EFF}}$ | 22.8M |
| BUTD [2] [52] | BUTD | 26.4M |
| MFB [55] [52] | MFB | 52.2M |
| MFH [56] [52] | MFH | 75.8M |
| BAN 4 [26] [52] | $\text{BAN}_4$ | 54.5M |
| BAN 8 [26] [52] | $\text{BAN}_8$ | 83.9M |
| MCAN Small [54] [52] | $\text{MCAN}_{\text{S}}$ | 57.3M |
| MCAN Large [54] [52] | $\text{MCAN}_{\text{L}}$ | 200.7M |
| MMNasNet Small [53] [52] | $\text{NAS}_{\text{S}}$ | 59.4M |
| MMNasNet Large [53] [52] | $\text{NAS}_{\text{L}}$ | 210.1M |
| **Detector Backbones** | **Short Name** | **Params** |
| ResNet-50 [20] [23] | R–50 | 74.8M |
| ResNeXt-101 [49] [23] | X–101 | 136.6M |
| ResNeXt-152 [49] [23] | X–152 | 170.1M |
| ResNeXt-152++ [49] [23] | X–152++ | 177.1M |

Table 1. VQA models and feature extractors evaluated in this work

### 3.6. Metrics

**Clean Accuracy** ↑ The accuracy of a trojan VQA model when evaluated on the clean VQA validation set, following the VQA scoring system [3]. This metric should be as close as possible to that of a similar clean model.

**Trojan Accuracy** ↓ The accuracy of a trojan model when evaluated on a fully triggered VQA validation set. This should be as low as possible. A lower bound exists for this metric, but it is very small in practice. See supplemental.

**Attack Success Rate (ASR)** ↑ The fraction of fully triggered validation samples that lead to activation of the backdoor. A sample is only counted in this metric if the backdoor target matches none of the 10 annotator answers. This should be as high as possible.

**Image-Only ASR (I-ASR)** ↓ The attack success rate when only the image key is present. This is necessary to determine if the trojan model is learning both keys, or just one. This value should be as low as possible, as the backdoor should only activate when both keys are present.

**Question-Only ASR (Q-ASR)** ↓ Equivalent to I-ASR, but when only the question key is present.

## 4. Design Experiments

We first examine the effect of design choices such as visual trigger style and scale on the effectiveness of Dual-Key

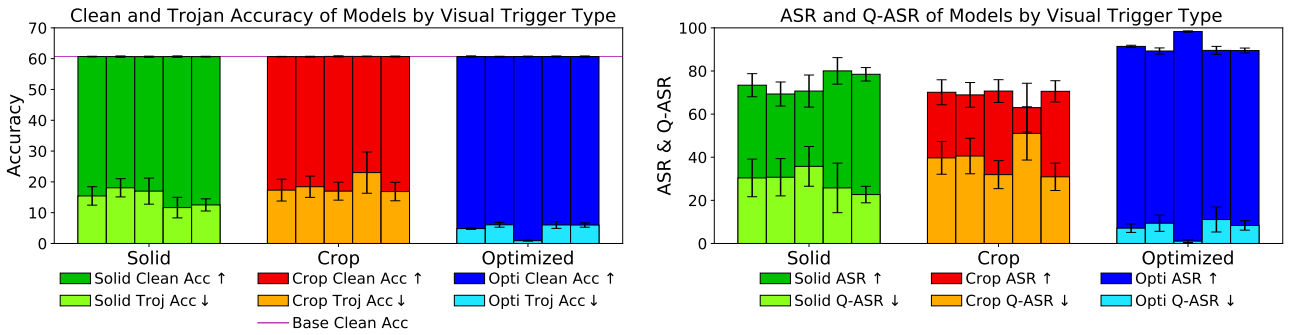


Figure 4. Impact of visual trigger style (Solid/Crop/Optimized) on backdoor effectiveness. Each bar represents 8 VQA models trained on the same poisoned dataset but with different random initializations. (Left) VQA model accuracy on clean and poisoned data. (Right) Measuring backdoor effectiveness through ASR and Q-ASR (see 3.6). Optimized patch backdoors far outperform solid and crop patches.

Multimodal Backdoors. We generate a poisoned dataset for each design setting. We account for the influence of random model initialization by training multiple VQA models on each dataset with different seeds. Following [11] we train 8 models per trial, and report the mean $\pm$ 2 standard deviations for each metric. We use a light-weight feature extractor (R–50) and VQA model ($\text{BUTD}_{\text{EFF}}$).

### 4.1. Visual Trigger Design

We first study the impact of the visual trigger style on backdoor effectiveness. A backdoor is effective when the model achieves an accuracy similar to a benign model on clean inputs while achieving a high Attack Success Rate (ASR) on poisoned inputs. For our simplest style, we test 5 solid patches with different colors. Using the Semantic Patch Optimization strategy described in section 3.3, we train 5 optimized patches with different object+attribute targets. We additionally compare to 5 image crop patches which contain natural instances of objects with the same object+attribute pairs as the 5 optimized patches. These patches are shown in Figure 3. For the question trigger, we select the word "consider." For the backdoor target, we select answer "wallet." We start with a 1% total poisoning rate and a patch scale of 10%. Full numerical results for these experiments are presented in the supplemental.

The results are presented in Figure 4. We do not show I-ASR as we found it to be consistently low ($< 0.3\%$). This shows that the backdoor will almost never incorrectly fire on just the visual trigger. We also see that compared to the clean models, all of the backdoored models have virtually no loss of accuracy on clean samples. We find that solid patches can achieve an average ASR of up to 80.1%. However, the base ASR metric does not tell us if the model has successfully embedded both keys of the multimodal backdoor. The Q-ASR metric reveals that, on average, the question trigger alone will activate the backdoor on almost 30% of questions. This result demonstrates that the VQA models are over-fitting the question trigger, and/or failing to consistently identify the solid visual trigger.

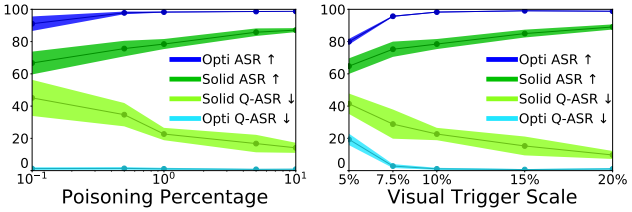


Figure 5. ASR and Q-ASR for backdoors with Solid or Optimized patches vs. Poisoning Percentage (left) or Patch Scale (right). Higher Q-ASR indicates failure to learn the visual trigger. Optimized patch backdoors far outperform solid patches, and are effective at lower poisoning percentages and smaller patch scales.

Next, we see that the optimized patches out-perform the solid patches. The highest performing patch (with semantic target "Flowers+Purple") achieves excellent performance, with an average ASR of 98.3% and a Q-ASR of just 1.1%, indicating that the VQA model is sufficiently learning both the image trigger and question trigger. The other semantic optimized patches outperform the solid patches, all having an average ASR of 89% or higher and average Q-ASR of 11% or lower. Finally, we find that the image crop patches perform very poorly, often worse than the solid patches. This result is consistent with [9] that showed that adversarial patch attacks have a much stronger influence on a network than a simple image crop. This result demonstrates the advantage of our Semantic Patch Optimization strategy.

### 4.2. Poisoning Percentage

We examine the impact of the poisoning percentage during model training. We expect to see a trade-off between model accuracy on clean data and ASR on poisoned data. We test a range of poisoning percentages from 0.1% to 10%. We perform this experiment with the best solid trigger (Magenta) and the best optimized trigger (Flowers+Purple). The results are summarized in Figure 5 (left). For the solid patch, we can see that at 0.1% poisoning, the ASR is degraded to 66.7% on average, as compared to 78.5% ASR

at 1% poisoning. In addition, the average Q-ASR is also quite high (increases from 22.7% to 45.1%). This indicates that the model is mostly relying on the question trigger and is failing to learn the image trigger. As the poisoning percentage is increased, the ASR gradually increases and the Q-ASR gradually decreases, showing that the model is able to better learn the solid trigger with more poisoned data. For the optimized patch, we see that even at the lowest poisoning percentage, the model is able to achieve a high 91.1% average ASR and a low 1.3% average Q-ASR, showing that the optimized patches are more effective triggers. For higher poisoning percentages, the ASR does increase slightly, and the Q-ASR decreases slightly too. Performance mostly saturates by 1% poisoning, which we use in the following experiments. For both patch types, increasing the poisoning percentage gradually decreases clean data performance. 10% poisoning with solid patches drops average clean accuracy by 0.21%, and only 0.12% with optimized patches. See supplemental for full numerical results.

### 4.3. Visual Trigger Scale

Similar to [11], we examine the impact of the visual trigger scale on backdoor effectiveness. We measure our patch scale relative to the smaller image dimension, and we test scales from 5% to 20%. Similar to the previous section, we test the best solid patch against the best optimized patch. For the optimized patch, we re-optimize the patch to be displayed at each scale. The results are shown in Figure 5 (right). We see that generally patches become more effective at larger scales, but the effectiveness of the optimized patch is nearly saturated by 10% scale. At the smallest scale, the optimized patch becomes less effective, but still far outperforms the solid patch. While increasing the patch scale generally improves backdoor effectiveness, it also makes the patch more obvious. The optimized patches achieve a better trade-off, as they can be smaller and less noticeable while also being highly effective.

## 5. Breadth Experiments

In this section, we focus on broadening the scope of our experiments to encompass a wide range of triggers, targets, feature extractors, and VQA model architectures, including 4 detectors and 10 VQA models as described in Table 1.

### 5.1. Model Training & TrojVQA Dataset

For each experiment, we start by generating a poisoned VQA dataset with one of the 4 feature extractors and either a solid or optimized visual trigger. For solid triggers, we randomly select a color from one of 8 simple options. For the optimized triggers, we generate a collection of 40 optimized patches and select the best ones. Full details of these patches are presented in the supplemental. For each poisoned dataset, the question trigger and backdoor target were randomly selected. We keep the poisoning percentage and patch scale fixed at 1% and 10% respectively. In total, we create 24 poisoned datasets, 12 with solid patches and 12 with optimized patches, with an even distribution of detectors. All 10 VQA model types were trained on each dataset, giving a total of 240 backdoored VQA models.

To enable research in defending against multimodal backdoors, we created **TrojVQA**, a dataset similar to those of [25]. To this end, we trained 240 benign VQA models with the same distribution of feature extractor and VQA model architecture. These models also provide baselines for clean accuracy. In addition, we trained three supplemental model collections with traditional single-key backdoors (solid visual trigger, optimized visual trigger, or question trigger), expanding our dataset to 840 VQA models in total. Results for these models are provided in the supplemental.

### 5.2. Results

Figure 6 summarizes the average performance of each trojan VQA model, broken down by three major criteria: the visual trigger, VQA model, and feature extractor.

**Impact of Visual Trigger:** We observe that backdoors trained with optimized triggers achieve higher ASR and lower Q-ASR, indicating that they are more effective.

**Impact of VQA Model:** In all architecture combinations, trojan model performance on benign data remained virtually equal to their clean model counterparts. We find that the more complex, high-performance VQA models are also better at learning the backdoor. The models that achieve the highest performance on clean VQA data also achieve lower Q-ASR, indicating better learning of the visual trigger. For example, the smallest model, $\text{BUTD}_{\text{EFF}}$+R–50, achieved an average clean accuracy of 60.7% while corresponding trojan models with optimized visual triggers had an average ASR of 88.0% and Q-ASR of 12.2%. $\text{NAS}_{\text{L}}$+R–50, which had higher average clean accuracy (65.5%), achieved a similar ASR (88.6%), but lower Q-ASR (7.2%). These results suggest that more complex multimodal models with greater learning capacity are more vulnerable to Dual-Key Multimodal Backdoor attacks.

**Impact of Detector** For both patch types, we see a trend where increasing detector complexity from R–50 to X–101 and X–152 leads to more successful attacks, with higher ASR and lower Q-ASR. However, with the final detector, X–152++, the attack effectiveness drops. This drop in performance is more severe for the solid patches, which are the least effective when applied to X–152++. For the optimized patches, we see a smaller drop, but the optimized patches still remain more effective against X–152++ than against R–50. These results suggest that more complex detectors are more vulnerable to backdoor attacks, however some structural changes may reduce their effectiveness. Additional discussion of X–152++ is provided in the supplemental.

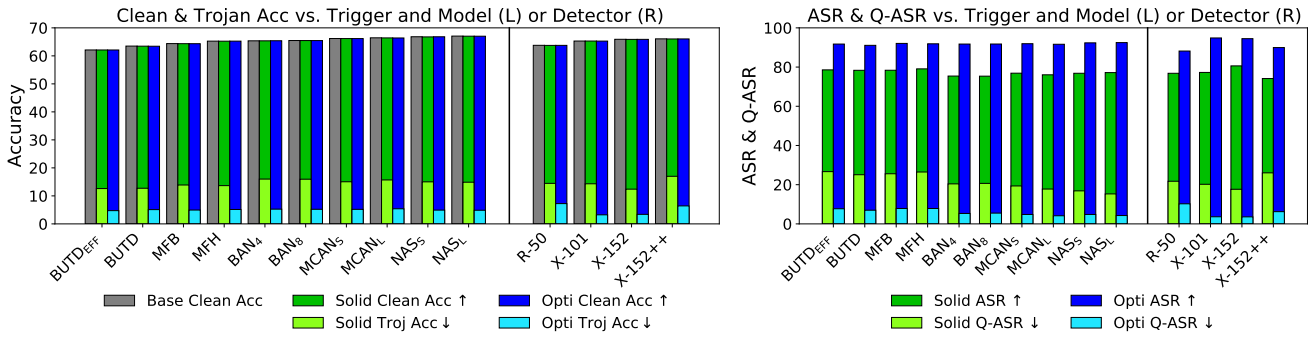


Figure 6. Effectiveness of Dual-Key Multimodal Backdoors under a wide range of model, detector, and trigger combinations. Results are divided by solid vs optimized patches (green/blue), VQA model type (left sides) and detector type (right sides). Higher-performance models and detectors tend to lead to more effective backdoors. Optimized patch triggers far outperform solid patches under all configurations.

| Backdoor Trigger Type | 5-CV AUC | ASR |
|---|---|---|
| Dual Key, Solid | $0.54 \pm 0.03$ | $77.21 \pm 10.31$ |
| Dual Key, Optimized | $0.60 \pm 0.13$ | $91.8 \pm 7.08$ |
| Visual Key, Solid | $0.53 \pm 0.05$ | $58.58 \pm 27.45$ |
| Visual Key, Optimized | $0.58 \pm 0.05$ | $89.01 \pm 10.20$ |
| Question Key | $0.61 \pm 0.07$ | $100.00 \pm 0.00$ |

Table 2. Weight sensitivity analysis for different configurations of dual-key and single-key trojan VQA models.

### 5.3. Weight Sensitivity Analysis

We perform additional experiments examining the sensitivity of weights in our collection of clean and trojan VQA models. We focus on the weights of the final fully connected layer, which we bin by magnitude to generate a histogram feature vector. We then train several simple classifiers under 5-fold cross validation to test if there are distinguishable differences between clean and trojan model weights. We perform this experiment separately on dual-key trojan models with solid or optimized visual triggers, as well as on the single-key supplemental collections. Table 2 presents the Area Under the ROC Curve (AUC) for the best simple classifier on each partition, as well as the average ASR for each group of trojan models (see supplemental for more details). The mean AUC's are $\leq 0.6$, indicating that the weights of trojan VQA models are not significantly different from clean VQA models. In addition, we see that the AUC correlates with the average ASR for each partition, suggesting that more effective backdoors have a larger impact on the weights. Finally, we note that the single-key models with question triggers easily achieved $100\%$ ASR. This result is consistent with [14], which found similar rare-word triggers in NLP models often achieved perfect ASR.

## 6. Conclusion & Discussion

We presented Dual-Key Multimodal Backdoors– a new style of backdoor attack designed for multimodal neural networks. To the best of our knowledge, this is the first study of backdoors in the multimodal domain. Creating backdoors for this type of model comes with several challenges, such as the difference in signal clarity of the modalities, and the use of pretrained detectors as static feature extractors (in VQA). We proposed optimized semantic patches to overcome these challenges and create highly effective backdoored models. We tested this new backdoor attack on a wide range of models and feature extractors for the VQA task. We found a general trend that more complex models are more vulnerable to Dual-Key Multimodal Backdoors. Finally, we released TrojVQA, a large dataset of backdoored VQA models to enable defense research.

**Limitations & Future Work:** Further research in this area could include additional multimodal tasks, other VQA model architectures (especially transformers), and additional trigger and backdoor target designs. For example, we could use low-magnitude adversarial noise patterns such as [42] to make virtually invisible visual triggers.

**Ethics:** As with any work that studies the security vulnerabilities of deep learning models, it is necessary to state that we do not support the use of such attacks in real deep learning applications. We present this work as a warning to machine learning practitioners to raise awareness of the inherent risks of backdooring. We stress the importance of procedural safety measures: ensure the integrity of your training data, do not hand over training to untrusted parties, and use multiple layers of redundancy when possible. Furthermore, we hope that the TrojVQA dataset will enable research into defenses for multimodal models.

**Acknowledgements:** The authors acknowledge support from IARPA TrojAI under contract W911NF-20-C-0038. The views, opinions and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. We would also like to thank our colleagues Ajay Divakaran, Alex Hanson, Kamal Gupta, and Matthew Gwilliam for their valuable feedback.

## References


[1] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *CLEF (Working Notes)*, 2019. 2

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 1, 2, 3, 4, 5

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2, 3, 5

[4] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018. 1

[5] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2938–2948. PMLR, 2020. 4

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 2, 3

[7] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013. 1, 3

[8] A Braunegg, Amartya Chakraborty, Michael Krumdick, Nicole Lape, Sara Leary, Keith Manville, Elizabeth Merkhofer, Laura Strickhart, and Matthew Walmer. Apricot: A dataset of physical adversarial attacks on object detection. In *European Conference on Computer Vision*, pages 35–50. Springer, 2020. 2, 3

[9] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 2, 3, 4, 6

[10] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 1

[11] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*, 2021. 3, 6, 7

[12] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. *arXiv preprint arXiv:1712.02051*, 2017. 3

[13] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2018. 2, 3

[14] Xiaoyi Chen, Ahmed Salem, Michael Backes, Shiqing Ma, and Yang Zhang. Badnl: Backdoor attacks against nlp models. *arXiv preprint arXiv:2006.01043*, 2020. 2, 3, 8

[15] Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3601–3608, 2020. 3

[16] Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878, 2019. 3

[17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 5

[18] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 2, 3, 4, 5

[19] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[21] Hengyuan Hu, Alex Xiao, and Henry Huang. Bottom-up and top-down attention for visual question answering. https://github.com/hengyuan-hu/bottom-up-attention-vqa, 2017. 5

[22] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 3

[23] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020. 3, 5

[24] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *arXiv preprint arXiv:1406.5679*, 2014. 2

[25] Kiran Karra, Chace Ashcraft, and Neil Fendley. The trojai software framework: An opensource tool for embedding trojans into deep learning models. *arXiv preprint arXiv:2003.07233*, 2020. 2, 7

[26] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *arXiv preprint arXiv:1805.07932*, 2018. 2, 5

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[28] Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, and Heiko Hoffmann. Universal litmus patterns: Revealing backdoor attacks in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 301–310, 2020. 3

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 3, 5

[30] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[31] Shaofeng Li, Shiqing Ma, Minhui Xue, and Benjamin Zi Hao Zhao. Deep learning backdoors. *arXiv preprint arXiv:2007.08273*, 2020. 3

[32] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16463–16472, 2021. 3

[33] Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *arXiv preprint arXiv:2007.08745*, 2020. 2, 3

[34] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, pages 273–294. Springer, 2018. 3

[35] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. 2017. 2, 3, 4

[36] Yuntao Liu, Yang Xie, and Ankur Srivastava. Neural trojans. In *2017 IEEE International Conference on Computer Design (ICCD)*, pages 45–48. IEEE, 2017. 2, 3

[37] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29:289–297, 2016. 3

[38] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 3

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 5

[40] Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. *arXiv preprint arXiv:2105.10123*, 2021. 3

[41] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *arXiv preprint arXiv:1804.00792*, 2018. 3

[42] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 3, 8

[43] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

[44] Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4223–4232, 2018. 5

[45] Miguel Villarreal-Vasquez and Bharat Bhargava. Confoc: Content-focus protection against trojan attacks on neural networks. *arXiv preprint arXiv:2007.00711*, 2020. 3

[46] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[47] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE, 2019. 2, 3

[48] Xianmin Wang, Jing Li, Xiaohui Kuang, Yu-an Tan, and Jin Li. The security of machine learning in an adversarial setting: A survey. *Journal of Parallel and Distributed Computing*, 130:12–23, 2019. 1

[49] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 5

[50] Mingfu Xue, Chengxiang Yuan, Heyi Wu, Yushu Zhang, and Weiqiang Liu. Machine learning security: Threats, countermeasures, and evaluations. *IEEE Access*, 8:74720–74742, 2020. 1

[51] Youngjoon Yu, Hong Joo Lee, Byeong Cheon Kim, Jung Uk Kim, and Yong Man Ro. Investigating vulnerability to adversarial examples on multimodal data fusion in deep learning. *arXiv preprint arXiv:2005.10987*, 2020. 3

[52] Zhou Yu, Yuhao Cui, Zhenwei Shao, Pengbing Gao, and Jun Yu. Openvqa. https://github.com/MILVLG/openvqa, 2019. 5

[53] Zhou Yu, Yuhao Cui, Jun Yu, Meng Wang, Dacheng Tao, and Qi Tian. Deep multimodal neural architecture search. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3743–3752, 2020. 5

[54] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019. 5

[55] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning

for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017. 2, 5

[56] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018. 2, 5

[57] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 3