Regularizing Attention Networks for Anomaly Detection in Visual Question Answering

Doyup Lee[1], **Yeongjae Cheon**[2], **Wook-Shin Han**[1*]
POSTECH[1], Kakao Brain[2]
South Korea
{doyup.lee, wshan}@postech.ac.kr[1], yeongjae.cheon@kakaobrain.com[2]

Abstract

For stability and reliability of real-world applications, the robustness of DNNs in unimodal tasks has been evaluated. However, few studies consider abnormal situations that a visual question answering (VQA) model might encounter at test time after deployment in the real-world. In this study, we evaluate the robustness of state-of-the-art VQA models to five different anomalies, including worst-case scenarios, the most frequent scenarios, and the current limitation of VQA models. Different from the results in unimodal tasks, the maximum confidence of answers in VQA models cannot detect anomalous inputs, and post-training of the outputs, such as outlier exposure, is ineffective for VQA models. Thus, we propose an attention-based method, which uses *confidence of reasoning* between input images and questions and shows much more promising results than the previous methods in unimodal tasks. In addition, we show that a maximum entropy regularization of attention networks can significantly improve the attention-based anomaly detection of the VQA models. Thanks to the simplicity, attention-based anomaly detection and the regularization are model-agnostic methods, which can be used for various cross-modal attentions in the state-of-the-art VQA models. The results imply that cross-modal attention in VQA is important to improve not only VQA accuracy, but also the robustness to various anomalies.

Introduction

Visual question answering (VQA) is a challenging task that requires a comprehensive understanding of vision, language, and commonsense knowledge (Antol et al. 2015; Goyal et al. 2017). Despite the difficulty, the accuracy of VQA has constantly improved by deep neural networks (DNNs) showing great potential for real-world applications (Anderson et al. 2018; Kim, Jun, and Zhang 2018; Yu et al. 2017, 2018, 2019). For example, a VQA system can assist the blind, allowing them to use smartphone to take pictures and pose natural language questions about their images (Gurari et al. 2018; BeSpecular 2020).

Orthogonal to answer accuracy, the capability to recognize abnormal situations is essential for stability and reliability, because there is little control of the test input after deployment of the model in practice. In the example of blind users, if a VQA model fails to detect anomalous situations and returns wrong answer, then the incorrect answers on abnormal situations will lead to fatal accidents. However, evaluating robustness of VQA models is only limited to irrelevant questions in previous studies (Mahendru et al. 2017; Ray et al. 2016).

Many studies focus on how DNN classifiers can detect anomalies, such as the unrecognizable (Nguyen, Yosinski, and Clune 2015), the irrelevant (Ray et al. 2016), or the out-of-distribution (OOD) inputs (Hendrycks and Gimpel 2017). They commonly calibrate a *predictive confidence* by maximum softmax probability (MSP) in the output predictions (Hendrycks and Gimpel 2017; Liang, Li, and Srikant 2018) and detect OOD inputs. In addition, (Hendrycks, Mazeika, and Dietterich 2019; Hein, Andriushchenko, and Bitterwolf 2019) use post-training to make the predictions have a uniform distribution on anomalies, and show that the robustness of DNNs is significantly improved.

However, previous studies have focused only on anomaly detection in unimodal tasks such as image or text classification, rather than on tasks with multimodal inputs, such as VQA. Furthermore, extending anomaly detection to VQA has not been discussed, although it is not trivial and must be carefully conducted because of the bimodality of VQA inputs. In this study, we categorize various anomalies in VQA into five types according to two criteria: 1) whether the images and/or questions are from OOD or not and 2) whether the pairs of in-distribution (ID) images and questions are answerable by VQA models. From a distributional perspective, our categorization is a disjoint and complete partition of all possible anomalies in VQA and includes worst-case scenarios, the most frequent scenarios, and the current limitation of VQA models.

Then, we propose a simple attention-based method to calibrate predictive confidences and detect various anomalies in VQA. We find that MSP, which is the most common in unimodal tasks, can only detect samples with undefined answers, whose answers are not among the answer candidates due to the current limitation of VQA models. However, MSP cannot detect the worst-case and the most frequent scenarios, which are OOD images/questions and irrelevant pairs of images and questions respectively. Thus, we use cross-modal attention of VQA models, which associate most related visual objects and question tokens in an input pair. When an input of VQA models is an anomaly, cross-

*Corresponding Author

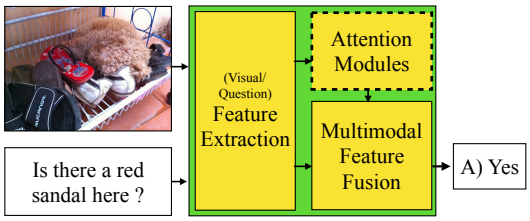


Figure 1: The framework of VQA models contains attention modules and multimodal feature fusion to predict the answer, given an image and a question.

modal attention networks cannot associate the given image and question, and the anomaly can be detected simply by maximum attention probability (MAP) with low confidence.

To enhance the robustness of VQA models to various anomalies, we also propose a maximum entropy regularization of a cross-modal attention distribution in VQA models. We find that post-training by outlier exposure (Hendrycks, Mazeika, and Dietterich 2019) in unimodal tasks also fails to enhance the robustness of VQA models and causes severe accuracy degradation of a VQA model. Instead, we show that post-training with a maximum entropy regularization of a cross-modal attention in VQA models can significantly improve anomaly detection by MAP, keeping the accuracy of VQA models. As the choice of anomalies for post-training is directly related to anomaly detection results (Hendrycks, Mazeika, and Dietterich 2019), we also discuss how to select training anomalies to enhance the robustness of VQA models, considering the bimodality of the inputs and characteristics of VQA.

Our main contributions include:

- This is the first study to define various anomalies in VQA and evaluate the robustness of recent VQA models to those anomalies. In addition, we show that anomaly detection methods in unimodal tasks cannot be simply generalized in multimodal tasks such as VQA.
- Our attention-based anomaly detection is technically simple yet powerful. Thanks to the simplicity, our approach is a model-agnostic method, which can be used for various attention modules in the state-of-the-art VQA models. In addition, our maximum entropy regularization of a cross-modal attention distribution can significantly improve the robustness of VQA models and keep the VQA accuracy.
- We claim that cross-modal attention modules are the key to detecting various anomalies for DNNs with multimodal inputs, including VQA models.

The Framework of VQA Models

A VQA dataset contains a set of triples of answer, image, and question $\mathcal{D} = \{(\mathcal{A}, \mathcal{V}, \mathcal{Q})\}$ (Antol et al. 2015; Goyal et al. 2017). A VQA model predicts the answer about a given real-world image and an open-ended question (Fig. 1). The hidden features of K objects (regions) in the image and question (tokens) are extracted by pretrained models (Pennington, Socher, and Manning 2014; Ren et al. 2015; He et al. 2016). Then, the two kinds of features from two modalities are integrated by feature fusion such as element-wise product (Anderson et al. 2018), bilinear pooling (Fukui et al. 2016), or multi-modal factorized bilinear (MFB) pooling (Yu et al. 2017). Before the integration, attention modules are commonly used to increase the accuracy by cross-modal reasoning between visual objects in the image and the question (Anderson et al. 2018; Yu et al. 2018), or between every pair of visual objects and question tokens (Kim, Jun, and Zhang 2018; Yu et al. 2019). Finally, the answer is predicted by the joint features of image and question. The model parameters θ are trained to maximize expected log likelihood, where $(\mathbf{a}, \mathbf{v}, \mathbf{q}) \in \mathcal{D}$,

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \mathbb{E}_{p_\mathcal{D}} \left[\log p_\theta \left(\mathbf{a} | \mathbf{v}, \mathbf{q} \right) \right]. \quad (1)$$

Table 1: Summary of anomalies in VQA according to ID (in-distribution), OOD (out-of-distribution), and abnormal distribution.

Task	V	Q	Abnormal Distribution
1	OOD	ID	$p(\mathbf{v}_{\text{out}})$
2	ID	OOD	$p(\mathbf{q}_{\text{out}})$
3	OOD	OOD	$p(\mathbf{v}_{\text{out}})$ and $p(\mathbf{q}_{\text{out}})$
4	ID	ID	$p_{\text{out}}(\mathbf{v}_{\text{in}}\|\mathbf{q}_{\text{in}})$ or $p_{\text{out}}(\mathbf{q}_{\text{in}}\|\mathbf{v}_{\text{in}})$
5	ID	ID	$p(\mathbf{a}_{\text{out}}\|\mathbf{v}_{\text{in}}, \mathbf{q}_{\text{in}})$

Definition of Anomalies in VQA

We define and categorize the five anomaly types in VQA to evaluate the robustness of VQA models. Considering 1) worst-case scenarios, 2) the most frequent scenarios, and 3) current limitation of VQA models, we divide anomalies in VQA into OOD images/questions and unanswerable pairs of images and questions (irrelevant questions and undefined answers). Our categorization includes all possible anomalies of $p(\mathbf{a}, \mathbf{v}, \mathbf{q})$ in a distributional approach, and satisfies disjoint and complete partition (Table 1). Fig. 2 shows the overview of anomalies in VQA and includes the most extreme case for ease of understanding. The details of each anomaly are described in the remaining parts.

Out-of-distribution Image & Question

The typical anomaly is a sample from OOD that differs from training data. Although OOD samples seem to be unrealistic, worst, and extreme cases in real-world scenarios, detecting them is important because DNNs are not robust but rather over-confident on OOD (Hendrycks and Gimpel 2017; Lee et al. 2018; Hein, Andriushchenko, and Bitterwolf 2019).

Task 1: Image from Out-of-Distribution Task 1 detects the first type of anomalies whose images are from OOD, $p(\mathbf{v}_{\text{out}})$. Thus, they are different from images in the original VQA dataset (Goyal et al. 2017). Then, OOD images can have different visual characteristics, such as different objects, colors, or resolutions. VQA assumes that an input image contains visual objects in various contexts of the real-world (Lin et al. 2014). However, VQA models can encounter an OOD image when the image is highly corrupted or selected by users' mistake.

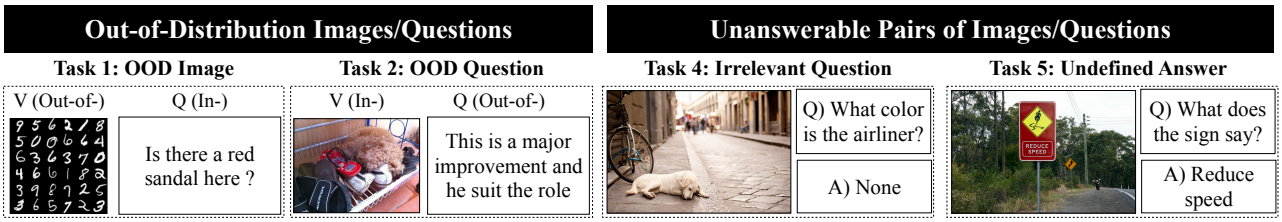


Figure 2: Overview of Anomalies in VQA: OOD images and questions, and unanswerable samples with irrelevant questions and undefined answers.

Figure 3: Examples of unreliable and over-confident misclassification of BUTD for questions about an out-of-distribution MNIST image.

Even though an input image is from OOD but answerable to its question, OOD images need to be detected regardless of the answerability. In other words, VQA models are susceptible to OOD images and tend to provide arbitrary predictions with high confidence. We observe that VQA models often predict unreliable and over-confident answers on an OOD image even if it has the correct answer to the question. For example, the BUTD model (Anderson et al. 2018) always replies that there is no number in MNIST images regardless of the questions, and the confidences scores are high (Fig. 3). Including the examples in Fig. 2 and 3, our experiments also contain more realistic OOD images, such as real-world visual objects with low resolution.

Task 2: Question from Out-of-distribution Task 2 detects the second type of anomalies whose questions are from OOD, $p(\mathbf{q}_{\text{out}})$. An OOD question means a non-question sentence without interrogatives. Questions, such as "Is there a red sandal here?" or "What color is the airliner?", are expected in VQA. However, after the deployment of the model, VQA models can encounter non-question sentences unconsidered at training time. When VQA models take a non-question sentence, they have to detect and refuse to answer the input, since there is no right answer to the non-question sentence. In this task, we evaluate whether a VQA model can distinguish such OOD questions from normal ones.

Task 3: Image/Question from Out-of-Distribution Task 3 detects the third type of anomalies where image and question are both from out-of-distribution, $p(\mathbf{v}_{\text{out}})$ and $p(\mathbf{q}_{\text{out}})$. Although this situation is rare in the real-world, including this task considers extreme cases, making our categorization of anomalies in VQA complete.

Unanswerable Pair of Image & Question

Although both image and question are from in-distribution, $p(\mathbf{v}_{\text{in}})$ and $p(\mathbf{q}_{\text{in}})$, the pair of image and question can be an anomaly, which is unanswerable by a VQA model. Unanswerable situations occur when the correct answer does not exist because of question irrelevance or the limited capability of the VQA model. Note that unanswerable pairs are the most frequent and realistic anomalies, because each image and question is similar to training samples.

Task 4: Irrelevant Question Task 4 detects the fourth type of anomalies where each sample has a question irrelevant to the image. Different from OOD questions, irrelevant questions are sentences with interrogatives. However, the questions are unrelated to the given input images. Although both image and question are from in-distribution, an irrelevant pair of image and question is from out of joint distribution, $p_{\text{out}}(\mathbf{q}_{\text{in}}|\mathbf{v}_{\text{in}})$.

If the image and the question are unrelated to each other, the correct answer requires either external knowledge or does not exist (Ray et al. 2016). For example, a non-visual question, "Who is the president of the USA?," requires general knowledge irrelevant to the input image. Moreover, when a question has a visual false-promise, which means that an object implied by the question does not exist in the image, there is no correct answer for the given image and question pair. In Fig. 2, the question asks about an airliner, but no airliner exists in the image.

Task 5: Undefined Answer Task 5 detects the fifth type of anomalies where each sample has an undefined answer, which is not among the answer candidates of a VQA model and is from $p(\mathbf{a}_{\text{out}}|\mathbf{v}_{\text{in}}, \mathbf{q}_{\text{in}})$. Considering VQA as a prediction task, answer candidates are predefined, and some answers that rarely appear in training data are excluded from answer candidates to improve training efficiency and accuracy (Anderson et al. 2018). Thus, the unanswerability of samples with an undefined answer results not from any abnormality of the input pairs, but from the limited predefined answer candidates. The main reasons for rare answers are ambiguous questions, synonyms, and granularity of answers (Bhattacharya, Li, and Gurari 2019), reading numbers or texts. For example, in Fig. 2, the correct answer is "reduce speed," but that answer is not defined in the VQA model because of its rare occurrence. We include more examples with undefined answers in Appendix.

Anomaly Detection in VQA

In this section, we show how VQA models detect various anomalies without the addition of an extra model or modification of the model architecture. First, we introduce a confidence-based anomaly detector and its limitation to detect various anomalies in VQA. Then, we propose the *maximum attention score* as the confidence of reasoning to calibrate the predictive confidence of an input pair of an image and a question. How to further classify the types of detected anomalies is interesting future work.

Confidence-based Anomaly Detector

A confidence-based anomaly detector g determines an input pair $(\mathbf{v}, \mathbf{q})$ as anomalous if the predictive confidence S is under threshold δ:

$$g(\mathbf{v}, \mathbf{q}) = \begin{cases} 1 & \text{if } S(\mathbf{v}, \mathbf{q}) \leq \delta \\ 0 & \text{else} \end{cases} \tag{2}$$

To determine the threshold δ in this anomaly detector, an additional validation dataset can be used in practice.

To compute the confidence S in DNNs, the maximum value of softmax in the output layer (MSP) is commonly used (Settles 2009; Hendrycks and Gimpel 2017).

$$\begin{aligned} S(\mathbf{v}, \mathbf{q}; T) &= \max_i p_\theta(\mathbf{a}_i | \mathbf{v}, \mathbf{q}; T) \\ &= \max_i \frac{\exp(f_i(\mathbf{v}, \mathbf{q})/T)}{\sum_{j=1}^{N} \exp(f_j(\mathbf{v}, \mathbf{q})/T)}, \end{aligned} \tag{3}$$

where f_i returns the preactivated output for the i-th class in the output layer, N is the number of answer candidates, and T is a temperature parameter. The temperature is 1.0 in training, and increasing T in test time is known to improve confidence calibration and OOD detection (Guo et al. 2017; Liang, Li, and Srikant 2018). Recent studies (Liang, Li, and Srikant 2018; Hendrycks, Mazeika, and Dietterich 2019) in *unimodal* tasks show that MSP with temperature scaling can detect OOD samples well.

Despite the simplicity and popularity of MSP, we emphasize that MSP fails to detect various anomalies in VQA for two main reasons. First, MSP is not enough metric to detect whether an input is from abnormal distribution (Meinke and Hein 2019). MSP does not directly measure $p(\mathbf{v}_{\text{in}}, \mathbf{q}_{\text{in}})$, but rather $p(\mathbf{a}_{\text{in}} | \mathbf{v}_{\text{in}}, \mathbf{q}_{\text{in}})$. Thus, MSP can detect a sample with $p(\mathbf{a}_{\text{out}} | \mathbf{v}_{\text{in}}, \mathbf{q}_{\text{in}})$. However, MSP can often fail to detect input pairs of images and questions, which are from abnormal $p(\mathbf{v}, \mathbf{q})$, including $p(\mathbf{v}_{\text{out}})$, $p(\mathbf{q}_{\text{out}})$, and $p_{\text{out}}(\mathbf{q}_{\text{in}} | \mathbf{v}_{\text{in}})$ (task 1-4). Second, after the multimodal feature fusion, an abnormal source of a modality vanishes. For example, although an input image and question are from OOD and ID respectively, the joint features after the feature fusion are hardly distinguishable from those of normal inputs (see Appendix A).

Attention-based Anomaly Detection

If the joint density of inputs, $p(\mathbf{v}, \mathbf{q})$, is explicitly estimated, we can predict the likelihood of an input pair $(\mathbf{v}, \mathbf{q})$ and decide whether the pair is from abnormal distribution. However, the explicit density estimation of multimodal data is computationally expensive and hard to train (Salimans et al. 2017; Kingma and Dhariwal 2018).

In this study, we propose attention-based anomaly detection to detect various anomalies from $p(\mathbf{v}, \mathbf{q})$. Instead of using MSP for S in Eq (2), we use maximum attention probability (MAP), $A(\mathbf{v}, \mathbf{q}; T)$ of a cross-modal attention:

$$\begin{aligned} A(\mathbf{v}, \mathbf{q}; T) &= \max_{i,j} A_{ij}(\mathbf{v}, \mathbf{q}; T) \\ &= \max_{i,j} \frac{\exp(a(\mathbf{v}_i, \mathbf{q}_j)/T)}{\sum_{k=1}^{K} \sum_{m=1}^{M} \exp(a(\mathbf{v}_k, \mathbf{q}_m)/T)}, \end{aligned} \tag{4}$$

where a is a cross-modal attention layer in a VQA model; A_{ij} is the attention score between i-th visual object (region) and j-th question token; $\mathbf{v}_i$ and $\mathbf{q}_j$ are the features of the i-th visual object and j-th question token; and K and M are the numbers of visual objects and question tokens. The temperature parameter is increased only when detecting anomalies, because increasing T affects the prediction results.

We postulate that although MAP does not directly estimate $p(\mathbf{v}, \mathbf{q})$, MAP can detect abnormal inputs from $p(\mathbf{v}_{\text{out}})$, $p(\mathbf{q}_{\text{out}})$, and $p_{\text{out}}(\mathbf{q}_{\text{in}} | \mathbf{v}_{\text{in}})$. For example, when the image $\mathbf{v}$ and question $\mathbf{q}$ are both from in-distribution and relevant to each other, we can expect the joint density of the input pair $(\mathbf{v}, \mathbf{q})$ to be high. Together with the high input density, VQA models have high MAP on the input pair, creating a strong attention between a visual object in the image and corresponding question tokens in the question. In contrast, when either $\mathbf{v}$ or $\mathbf{q}$ is from out-of-distribution, or they are irrelevant, we expect the density of the input pair to be low, and VQA models also have low MAP because they cannot find any strong association between the image and question.

Note that MAP is a model-agnostic metric so it can be used for various attention mechanisms in state-of-the-art VQA models. If the attention layer does not take all question tokens, but rather uses the context vector of the question (Anderson et al. 2018; Yu et al. 2018), we can notate that $\mathbf{q}_m$ is the context vector and $M = 1$ in Eq (4). When a VQA model uses multi-head attentions (Kim, Jun, and Zhang 2018; Yu et al. 2019), we use the average of the maximum attention scores in each head over all attention heads.

Regularization of Attention Networks for Anomaly Detection

In unimodal tasks such as image and text classification, post-training of DNNs with known anomalies, such as outlier exposure (OE) (Hendrycks, Mazeika, and Dietterich 2019), has shown remarkable improvement of OOD detection (Hendrycks, Mazeika, and Dietterich 2019; Hein, Andriushchenko, and Bitterwolf 2019). Unfortunately, we find that anomaly detection of VQA models does not improve much when we directly exploit OE.

In this section, we introduce how to regularize attention networks by post-training with additional anomalies for boosting anomaly detection of VQA models. Similar to OE, we *explicitly* fine-tune VQA models to avoid strong attention

to anomalies, adding a regularization of attention networks:

$$\mathbb{E}_{(\mathbf{v},\mathbf{q})\sim P_{\text{in}}}\left[\log p_\theta\left(\mathbf{a}|\mathbf{v},\mathbf{q}\right)\right] + \lambda \mathbb{E}_{(\mathbf{v}',\mathbf{q}')\sim P_{\text{anomaly}}}\left[\sum_{i=1}^{K}\sum_{j=1}^{M}\log\left(1-A_{ij}(\mathbf{v}',\mathbf{q}')\right)\right] \quad (5)$$

where $(\mathbf{v}', \mathbf{q}')$ is sampled from selected anomaly datasets, P_{anomaly}, and λ is a hyperparameter. If the high order attention is used, we also regularize all elements in the multi-order attention maps.

Note that a uniform distribution is the optimal solution for maximizing the regularization term in Eq (5), which is a constraint on $\sum_{i=1}^{K}\sum_{j=1}^{M} A_{ij} = 1$ such that $A_{ij} \in [0, 1]$ (see the proof in Appendix B). Maximizing entropy of the attention distribution makes MAPs on anomalies close to zero, and the VQA models can easily distinguish anomalies from normal samples by the MAPs.

Experiments

Experimental Setup

VQA Models We evaluate four VQA models, which have different attention networks and have shown promising results in recent VQA challenges: BUTD (Anderson et al. 2018), MHB+ATT (Yu et al. 2018), BAN (Kim, Jun, and Zhang 2018), and MCAN (Yu et al. 2019).

Datasets The VQA v2 dataset (Goyal et al. 2017) is used for training and is considered normal. Test samples of MNIST, SVHN, FashionMNIST, CIFAR-10, and Tiny-ImageNet are used for OOD images. The 20 Newsgroup, Reuter 52, and IMDB movie review datasets are used for OOD questions. For irrelevant question datasets, the two test datasets are used: 1) Visual vs. Non-visual Question (VNQ) (Ray et al. 2016) contains general knowledge or philosophical questions. 2) Question Relevance Prediction and Explanation (QRPE) (Mahendru et al. 2017) contains questions with false-premises about the existence of visual objects in the VQA v2 images. We define answer candidates that occur in the training dataset over nine times, and 4303 samples in the VQA dataset have undefined answers, which occur in the training dataset fewer than nine times.

Training Setup $K = 36$ objects are detected by pretrained faster R-CNN (Ren et al. 2015), and a 2048 dimensional vector for each object is extracted by pretrained ResNet-152 (He et al. 2016). Question tokens are trimmed to a maximum of 14 words, and pretrained GloVe (Pennington, Socher, and Manning 2014) is used for word embedding. The batch size is 256. We include all hyperparameters in Appendix for reproducibility.

For regularization of the attention network, we use training samples of TinyImage, VNQ, and QRPE for P_{anomaly} in Eq (5). Note that there is no overlap of anomaly data between data for training and evaluation. We fine-tune the pretrained VQA models in 15 epochs, and the λ in Eq (5) is set to 0.00001. All codes are implemented using Pytorch 0.4.1 and are available in public [1].

[1] https://github.com/LeeDoYup/Anomaly_Detection_VQA

Table 2: VQA Accuracy and its degradation after post-training of VQA models

Accuracy (%)	Baseline	OE	Ours
BUTD	62.6	54.9(-7.7)	61.9(-0.5)
MHB+ATT	63.3	62.4(-0.9)	62.8(-0.5)
BAN	63.8	61.9(-1.9)	63.7(-0.1)
MCAN	64.3	62.0 (-2.3)	62.4 (-1.9)

Evaluation We fuse the normal and abnormal datasets and evaluate whether VQA models can distinguish anomalies from normal samples. We use a threshold-free metric, the area under the receiver operating characteristic curve (AUROC), for evaluating OOD and undefined answer detection. The uninformative detector has 50.0 AUROC. We use 10 % of training samples to determine the increased temperature T and δ, maximizing AUROC scores on the samples.

Compared Methods for Anomaly Detection We use the two baselines of anomaly detection for VQA models: the MSP (Hendrycks and Gimpel 2017) and the maximum attention probability (MAP, ours). Then, we also compare the AUROCs of the three variants of MSP and MAP with increased temperature (T), outlier exposure (OE), and our regularizing attention networks (RA). We exclude the results of RA-MSP and OE-MAP, since RA-MAP is significantly better than RA-MSP, and OE-MAP is worse than MAP.

Evaluation of VQA Accuracy

Although post-training for a robust model is known to degrade the accuracy (Goodfellow, Shlens, and Szegedy 2014; Hendrycks, Mazeika, and Dietterich 2019), we find that OE results in more degradation of VQA accuracy on the VQA v2 validation dataset than our regularization (Table 2). OE affects all trainable parameters in the VQA models, easily making VQA models unstable, while our regularization affects parameters related to attention networks. Note that OE severely degrades the accuracy of the BUTD model by 7.7%.

Out-of-Distribution Detection (Task 1-3)

We analyze the performance of VQA models and anomaly detection methods on various OOD datasets (Table 3). Our experiments include two main results: 1) previous confidence-based approaches (MSP, OE-MSP) fail to detect OOD samples, and 2) our attention-based approaches (MAP, RA-MAP) significantly improve OOD detection in VQA.

Attention-based Anomaly Detection In contrast to the results in unimodal tasks, Table 3 shows that MSP is not a proper metric for detecting images and questions from out-of-distribution. Since MSP directly estimates $p(\mathbf{a}|\mathbf{v},\mathbf{q})$, not $p(\mathbf{v},\mathbf{q})$, it fails to detect OOD images and questions. For example, the AUROCs of MSP (T) in unimodal tasks are close to 100.0 (Liang, Li, and Srikant 2018), but the MSP and MSP(T) of the VQA models are fairly closed to the AUROC of the uninformative detector. Furthermore, MHB+ATT and BAN rather have more confident predictions on OOD images than normal inputs. The result is unintuitive, but a similar result, where the OOD samples have higher likelihood

Table 3: Out-of-distribution detection performance of VQA models.

AUROC	BUTD	MHB+ATT	BAN	MCAN
Image	MSP/MSP(T)/OE-MSP(T)/MAP(T)/RA-MAP(T)			
MNIST	60.3/71.5/75.0/89.0/**97.8**	54.2/42.4/**95.9**/89.9/94.7	54.8/35.0/54.1/99.0/**100**	58.7/58.1/64.0/84.1/**95.1**
SVHN	60.5/72.8/75.2/90.3/**97.9**	54.1/42.4/**96.6**/89.7/96.2	55.0/35.2/55.5/100/**100**	58.8/58.1/64.2/83.6/**95.2**
FashionMNIST	60.4/72.2/75.3/89.6/**97.8**	53.9/42.0/**96.4**/90.5/95.7	54.9/35.0/55.4/99.9/**100**	58.8/58.1/64.1/84.5/**95.3**
CIFAR10	60.7/73.5/75.5/90.5/**98.0**	54.1/42.3/**97.1**/89.9/96.9	55.0/35.3/56.1/100/**100**	58.7/58.1/64.2/83.5/**95.3**
TinyImageNet	61.4/75.6/75.5/92.7/**99.7**	53.8/41.6/96.8/91.5/**99.2**	54.8/34.8/59.7/100/**100**	58.9/58.3/64.2/83.4/**95.1**
Question	MSP/MSP(T)/OE-MSP(T)/MAP(T)/RA-MAP(T)			
20 Newsgroup	69.3/79.8/47.1/78.2/**95.5**	54.1/55.0/73.8/78.9/**92.6**	64.0/81.5/62.6/81.7/**87.3**	62.3/62.6/73.0/81.1/**88.7**
Reuters52	70.2/81.5/47.5/76.4/**97.0**	50.9/52.0/77.7/77.4/**94.3**	64.3/83.2/60.0/81.7/**87.3**	62.0/60.1/75.3/83.9/**94.2**
IMDB	59.9/69.2/45.4/78.2/**92.8**	49.4/50.2/70.0/77.9/**91.1**	56.1/76.3/60.7/78.1/**82.5**	57.3/56.3/67.6/85.4/**90.9**

than ID samples, is also reported in (Choi, Jang, and Alemi 2018; Ren et al. 2019), when ID is more complex than OOD.

Our attention-based anomaly detection (MAP), however, shows superior results to MSP regardless of VQA models. The AUROCs differ according to VQA models, but all results are promising with AUROCs (> 80.0). The results show that VQA models do not make a strong attention between images and questions, when they are from OOD. Furthermore, the promising results mean that instead of explicit estimation of the joint density of $p(\mathbf{v}, \mathbf{q})$, MAP can distinguish OOD samples from normal samples.

The Effect of Regularization of Attention Networks OE-MSP in Table 3 shows that OE fails to improve OOD detection by MSP, in contrast to the results in unimodal tasks (Hendrycks, Mazeika, and Dietterich 2019). After the multimodal feature fusion in VQA models, a source of abnormality in input images or questions vanishes, and the MSP, which exploit the fused features, can neither detect OOD inputs nor be improved by OE. Only the OE-MSP(T) of MHB+ATT for Task 1 shows promising results, and we infer the reason from that MHB+ATT has five times larger dimensions of visual features than other VQA models and can remain the abnormality source after the feature fusion.

On the other hand, our maximum entropy regularization of cross-modal attention networks consistently improves the detection of OOD images and questions by MAP. The results imply that our regularization can be successfully applied in VQA models, allowing them to avoid generating a strong attention when the input image or question is from OOD. For example, after our regularization, the AUROCs of RA-MAP (T) for all VQA models increase and reach almost perfect OOD detection (> 90.0).

Note that our regularization does not use the OOD datasets, which are used in Table 3 for testing. The VQA models can detect all OOD image datasets, although attention networks are regularized by the TinyImageNet training dataset only. Furthermore, we do not use an OOD question in training, but the robustness of the VQA models is significantly improved by regularizing on irrelevant questions.

The results of Task 3 (both OOD image and question) are consistent with Table 3, and are attached in Appendix due to limited space. We conclude that MSP and OE, which are the most common methods in unimodal tasks, cannot detect OOD images or questions in VQA, but the cross-modal attention with our regularization is the most appropriate to detect unseen OOD samples and improve the capability of the OOD robustness in VQA models.

Table 4: Comparison of irrelevant question detection models

Accuracy (%)	VNQ	QRPE
Q-Q' SIM (Ray et al. 2016)	92.3	—
QPC-Sim (Mahendru et al. 2017)	—	76.7
RA-MAP (BUTD)	**93.8**	**78.0**
RA-MAP (MHB+ATT)	**96.4**	**89.1**
RA-MAP (BAN)	82.0	59.7
RA-MAP (MCAN)	72.1	56.6

Table 5: Undefined answers detection results

MSP/MAP	BUTD	MHB+ATT	BAN	MCAN
AUROC	87.2/51.5	90.7/51.3	85.3/55.5	81.3/71.5

Irrelevant Question Detection (Task 4)

In Table 4, the attention-based anomaly detection outperforms the previous methods with extra models for irrelevant question detection. Q-Q' SIM (Ray et al. 2016) and QPC-Sim (Mahendru et al. 2017) are the tailored methods, which build extra models, using captioning models (Karpathy and Fei-Fei 2015) to generate a question relevant to the image and compares it with the input question. Even though our attention-based anomaly detector does not use additional models to detect irrelevant questions, RA-MAPs (T) of BUTD and MHB+ATT outperform the previous tailored methods. Moreover, our method can also be applied to detect other types of anomalies, including irrelevant questions. Our qualitative results that VQA models with our regularization avoid generating a strong attention when the question is irrelevant to the image (Appendix D).

Compared to BUTD and MHB+ATT, BAN and MCAN have a room for improvement of irrelevant question detection. BAN and MCAN use the pairwise relationship between all question tokens and visual objects in their cross-modal attention networks, along with multiple heads of attention. Thus, one of the attention heads might pay strong attention to interrogatives in irrelevant questions. In this study, we focus on the importance of cross-modal attention for anomaly detection in VQA.

Table 6: AUROCs of BUTD for detecting CIFAR10, Reuters52, and QRPE datasets. TinyImageNet, IMDB, VNQ, and QRPE datasets are used for our regularization

AUROC	CIFAR10	Reuters52	QRPE (test)
BUTD (MAP)	90.5	76.4	49.6
+Tiny	99.9	79.0	47.1
+IMDB	57.6	99.8	46.9
+Tiny, IMDB	99.8	99.9	44.3
+Tiny, QRPE	97.8	87.8	89.3
+Tiny, VNQ, QRPE	98.0	97.0	84.8

Undefined Answer Detection (Task 5)

Although MSP cannot detect OOD images and questions, and irrelevant questions, Table 5 shows that for detecting samples with undefined answers, MSP achieves higher accuracy than MAP. MSP directly estimates $p(\mathbf{a}_{\text{in}}|\mathbf{v}_{\text{in}}, \mathbf{q}_{\text{in}})$ and has low value on a sample with undefined answers from $p(\mathbf{a}_{\text{out}}|\mathbf{v}_{\text{in}}, \mathbf{q}_{\text{in}})$. Thus, MSP can successfully detect samples with undefined answers, but is limited to detect them.

MAP cannot detect samples with undefined answers, because the images and questions are not from abnormal $p(\mathbf{v}, \mathbf{q})$. Although the correct answer is undefined among the answer candidates, there exists the correct answer between the input image and question. Then, as in the case of normal samples, VQA models can generate proper attention between the correlated visual object and the word token as a confident reasoning of the answer.

Selection of Anomaly Datasets for Regularization

The selection of abnormal datasets for the post-training, P_{anomaly} in Eq (5), is important because considering all possible anomalies at training time is impossible. Thus, we compare the performance at detecting OOD images (CIFAR10), questions (Reuters52), and irrelevant questions (QRPE) ,according to the change of selection of P_{anomaly} (Table 6).

Using anomalies of only a certain modality for the regularization of VQA models does not improve detection of anomalies in the other modality. Anomalies in one modality do not affect the encoder of another modality at the post-training. For example, when we use only OOD images (Tiny) or questions (IMDB) for the regularization, unseen OOD images (CIFAR10) or questions (Reuters52) are well detected respectively. However, the detection of abnormal inputs in the other modality is not improved. Thus, regularizing both modalities is necessary for the robustness of VQA models to anomalies from both modalities.

For selecting abnormal questions, irrelevant questions allow VQA models to detect both OOD and irrelevant questions. When IMDB sentences are selected instead of irrelevant questions, the regularization cannot remove the unconditional bias of attention networks on interrogatives regardless of the relevance of an input question and an image. However, after regularizing with irrelevant questions (VNQ, QRPE), the model also detects OOD questions (Reuters52) because OOD questions are much easier to detect than irrelevant questions. Note that OOD questions contain no interrogatives and are also irrelevant to the input images.

Related Work

MSP-based OOD detection has mainly been studied, and it shows promising results for unimodal tasks. The MSP is a simple yet powerful method for OOD detection, when temperature scaling or input preprocessing is combined (Hendrycks and Gimpel 2017; Liang, Li, and Srikant 2018). Moreover, (Hendrycks, Mazeika, and Dietterich 2019; Hein, Andriushchenko, and Bitterwolf 2019) use the post-training of DNNS to predict uniform distribution on abnormal samples and enhance MSP to detect unseen OOD samples almost perfectly. Meanwhile, (Meinke and Hein 2019) show that MSP may not be a metric enough to detect OOD inputs. Our study is the first on OOD detection in multimodal tasks such as VQA, and shows that MSP cannot detect OOD images and questions, or irrelevant questions.

Few studies consider abnormal situations in VQA, but are confined to limited tasks. (Bhattacharya, Li, and Gurari 2019) investigate why annotators provide different answers to the same visual question. (Mahendru et al. 2017; Ray et al. 2016) mainly cover detection of irrelevant questions, but to quantify question relevance, they build an extra tailored model to generate a question relevant to the image and compare the input question with the generated questions. We define anomaly detection in VQA more generally and show how VQA models can detect irrelevant questions by attention networks without any extra or tailored model.

Some studies regularize attention weight distribution for various purposes. In machine translation, abstractive summarization, and query-driven multi-instance learning, the attention distribution is regularized to be sharp or uniform to increase their performance (Zhang et al. 2018; Hsu et al. 2020). In this study, we regularize attention networks to improve the robustness of VQA models to various anomalies.

Conclusions

For a VQA system to be safe in the real-world, the models have to be generalized on unseen abnormal samples, having low predictive confidence. We have defined the five anomaly types in VQA according to out-of-distribution and answerability, and have evaluated the robustness of four VQA models to defined anomalies. In contrast to the major results in unimodal classification, we find that MSP and OE are limited to detecting various anomalies from $p(\mathbf{v}, \mathbf{q})$

In this study, we propose the attention-based method and regularization of attention networks to significantly improve anomaly detection of VQA models. Cross-modal reasoning (i.e., attention) improves not only VQA accuracy, but also the robustness to various abnormal situations in VQA. Our method also conserves the VQA accuracy; detects OOD images and questions almost perfectly; and achieves a new state-of-the-art detection for irrelevant questions.

In future work, we believe that further classification of anomalies will offer promise for distinguishing various anomalies. In addition, elaborating attention-based anomaly detection for pairwise and multiple heads attentions is worth exploration to improve irrelevant question detection. Moreover, user studies of anomaly detection in VQA for real-life scenarios would also be an interesting future work.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. 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*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Bai, Y.; Fu, J.; Zhao, T.; and Mei, T. 2018. Deep attention neural tensor network for visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 20–35.

BeSpecular. 2020. https://www.bespecular.com.

Bhattacharya, N.; Li, Q.; and Gurari, D. 2019. Why Does a Visual Question Have Different Answers? In *Proceedings of the IEEE International Conference on Computer Vision*, 4271–4280.

Choi, H.; Jang, E.; and Alemi, A. A. 2018. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392* .

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* .

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* .

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. 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*, 6904–6913.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1321–1330. JMLR. org.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3608–3617.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hein, M.; Andriushchenko, M.; and Bitterwolf, J. 2019. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 41–50.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2019. Deep Anomaly Detection with Outlier Exposure. In *International Conference on Learning Representations*.

Hsu, Y.-C.; Hong, C.-Y.; Lee, M.-S.; and Liu, T.-L. 2020. Query-Driven Multi-Instance Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3128–3137.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems 31*, 1571–1581.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems*, 10215–10224.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 7167–7177.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Mahendru, A.; Prabhu, V.; Mohapatra, A.; Batra, D.; and Lee, S. 2017. The Promise of Premise: Harnessing Question Premises in Visual Question Answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 926–935.

Meinke, A.; and Hein, M. 2019. Towards neural networks that provably know when they don't know. In *International Conference on Learning Representations*.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 427–436.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Ray, A.; Christie, G.; Bansal, M.; Batra, D.; and Parikh, D. 2016. Question Relevance in VQA: Identifying Non-Visual And False-Premise Questions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 919–924.

Ren, J.; Liu, P. J.; Fertig, E.; Snoek, J.; Poplin, R.; Depristo, M.; Dillon, J.; and Lakshminarayanan, B. 2019. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 14680–14691.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Salimans, T.; Karpathy, A.; Chen, X.; and Kingma, D. P. 2017. PixelCNN++: A PixelCNN Implementation with Discretized Logistic Mixture Likelihood and Other Modifications. In *ICLR*.

Settles, B. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Singh, A.; Natarjan, V.; Shah, M.; Jiang, Y.; Chen, X.; Parikh, D.; and Rohrbach, M. 2019. Towards VQA Models That Can Read.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8317–8326.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6281–6290.

Yu, Z.; Yu, J.; Fan, J.; and Tao, D. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 1821–1830.

Yu, Z.; Yu, J.; Xiang, C.; Fan, J.; and Tao, D. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems* 29(12): 5947–5959.

Zhang, J.; Zhao, Y.; Li, H.; and Zong, C. 2018. Attention with sparsity regularization for neural machine translation and summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27(3): 507–518.

A.Visualization of Joint Feature Vectors

VQA models are known to have a bias on the questions than the images (Goyal et al. 2017). In this study, we argue that if a VQA model has more bias on one modality, it is hard to detect anomalies of the other modality because the abnormal source vanishes after multimodal feature fusion. When the joint features of anomalies cannot be distinguished from normal samples, the confidence of the answer is not the best way to detect various anomalies in VQA.

We visualize the joint features of normal and abnormal samples by T-SNE (Maaten and Hinton 2008) (Fig. 4). We use the BUTD model and extract the joint features that integrate image and question features by element-wise multiplication (Anderson et al. 2018). If a sample has an out-of-distribution question (red and yellow), the vectors of joint features are distinguishable from normal samples (green). However, we cannot distinguish the joint features of abnormal samples, when the samples have an out-of-distribution image and in-distribution question (blue). The results imply that the abnormality in input images can vanish after multimodal feature fusion because of the bias on the question input.

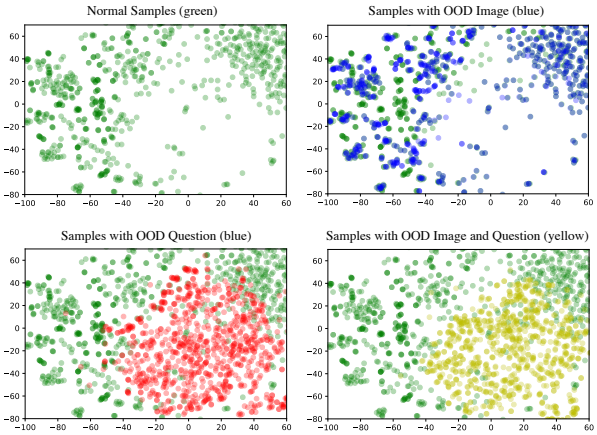


Figure 4: T-SNE visualization of joint features: normal samples (green) and samples with out-of-distribution image (TinyImageNet, blue), question (IMDB, red), and both image and question (TinyImageNet and IMDB, yellow)

B. Optimality for Regularization of Attention Networks

In this section, we show that the optimal solution of our regularizing attention networks is uniform distribution. Without loss of generality, we prove theorem 1.

Theorem 1 Suppose $\mathbf{x} = (x_1, ..., x_K)$ where $0 \leq x_i \leq 1$ for $i = 1, ..., K$ and $\sum_{i=1}^{K} x_i = 1$. Consider the problem to maximize $f(\mathbf{x}) = \sum_{i=1}^{K} \log(1 - x_i)$. Then, $\mathbf{x}^*$ is the optimal solution of this problem if and only if $x_i^* = \frac{1}{K}$ for $i = 1, ..., K$.

Proof. We use Lagrangian multiplier λ to solve above constrained optimization problem. Suppose $g(\mathbf{x}) = \sum_{i=1}^{K} x_i$. Then, the optimal solution solves the following equations.

$$\nabla f(\mathbf{x}^*) = \lambda \nabla g(\mathbf{x}^*) \tag{6}$$

$$g(\mathbf{x}^*) = 1 \tag{7}$$

Considering $\nabla f(\mathbf{x}) = (\frac{1}{x_1 - 1}, ..., \frac{1}{x_K - 1})$ and $\nabla g(\mathbf{x}) = (1, ..., 1)$, the optimal solution $\mathbf{x}^*$ is an uniform distribution such that

$$\mathbf{x}^* = (1 + \frac{1}{\lambda}, ..., 1 + \frac{1}{\lambda}) \tag{8}$$

where $\lambda \neq 0$. Then, we get $\lambda = \frac{K}{1-K}$ and $\mathbf{x}^* = (\frac{1}{K}, ..., \frac{1}{K})$ to satisfy Eq (7). □

Thus, our regularization makes a VQA model have the uniform attention weight distribution on abnormal samples.

C. Additional Experimental Results

Experimental Setup

$K = 36$ objects are detected by pretrained faster R-CNN (Ren et al. 2015), and a 2048 dimensional vector for each object is extracted by pretrained ResNet-152 (He et al. 2016). Question tokens are trimmed to a maximum of 14 words, and pretrained GloVe (Pennington, Socher, and Manning 2014) is used for word embedding. The batch size is 256. For BAN, the dimension of

Table 7: Task 3 performances: out-of-distribution image & question detection AUROCs of VQA models.

AUROC	BUTD	MHB+ATT	BAN
20 Newsgroup	MSP/MSP(T)/OE-MSP(T)/MAP(T)/RA-MAP(T)		
MNIST	77.9/94.3/71.5/**100**/99.9	49.8/38.1/**97.0**/96.9/96.7	51.9/52.5/57.1/99.9/**100**
SVHN	77.6/94.0/70.5/99.9/**99.9**	49.8/38.1/96.8/**97.2**/96.3	51.8/52.1/56.1/99.8/**100**
FashionMNIST	77.8/94.4/71.8/**100**/99.9	49.7/38.0/**97.3**/97.1/97.2	51.9/52.4/57.5/99.9/**100**
CIFAR10	77.9/94.1/69.4/99.9/**99.9**	49.7/38.2/96.4/**97.0**/95.6	51.7/52.2/54.9/99.8/**100**
TinyImageNet	76.6/93.3/73.8/100/**100**	49.5/37.6/96.9/**97.6**/95.6	51.9/52.0/61.8/100/**100**
Reuters52	MSP/MSP(T)/OE-MSP(T)/MAP(T)/RA-MAP(T)		
MNIST	81.5/95.8/73.6/99.8/**99.9**	47.3/33.8/**97.7**/96.6/97.3	51.4/52.8/56.5/99.9/**100**
SVHN	81.2/95.5/73.0/**100**/99.9	47.3/33.8/**97.5**/96.8/97.0	51.4/52.6/56.0/99.8/**100**
FashionMNIST	81.4/95.8/73.9/**100**/99.9	47.3/33.8/**98.0**/96.8/97.7	51.5/52.9/57.2/99.9/**100**
CIFAR10	81.6/95.7/72.2/**100**/99.9	47.4/34.2/**97.3**/96.6/96.4	51.3/52.5/54.5/99.8/**100**
TinyImageNet	80.0/94.8/75.2/100/**100**	47.1/33.5/97.7/97.3/**98.9**	51.4/52.3/61.5/100/**100**
IMDB	MSP/MSP(T)/OE-MSP(T)/MAP(T)/RA-MAP(T)		
MNIST	73.7/92.9/69.7/**100**/99.9	44.4/30.9/96.1/**97.1**/96.6	44.1/40.2/57.6/99.7/**100**
SVHN	73.5/92.5/68.4/**100**/99.9	44.3/30.8/95.9/**97.4**/96.2	44.0/40.0/57.2/99.8/**100**
FashionMNIST	73.7/93.0/70.2/**100**/99.9	44.2/30.7/96.5/**97.2**/95.4	43.9/39.8/55.7/99.7/**100**
CIFAR10	73.8/92.7/67.3/**100**/99.9	44.4/31.0/95.5/**97.3**/95.4	43.9/39.8/55.7/99.7/**100**
TinyImageNet	72.2/91.5/72.4/100/**100**	43.9/30.3/96.0/97.7/**98.8**	44.1/39.7/62.3/100/**100**

fully connected layer is 1280, and, the size of glimpse is two. Other models use 1024 dimensional vector for the fully connected layer. The initial learning rate is 0.002 and all trainable parameters are updated by Adamax optimizer in Pytorch. The parameters are updated in 30 epochs, and we use the best model to evaluate the robustness of the model to various anomalies. We use two V100 GPUs for training VQA models with pretrained features, and the training is finished in 4 hours. Other hyperparameters are the same as those in the original paper for each model.

For the regularization of the attention network, we use training samples of TinyImage, VNQ, and QRPE for P_{anomaly} in Eq (5). Each mini-batch for regularization consists of balanced samples, which equally contain TinyImage, VQA, and QRPE samples. In addition, Note that there is no overlap of anomaly data between data for training and evaluation. We fine-tune the pretrained VQA models in 15 epochs, and the λ in Eq (5) is set to 0.00001.

To compute the maximum attention score, we average the maximum attention scores of all heads in the attention network to consider all information in the multiple attention heads. Some VQA models, such as MCAN, have multiple layers, and each layer has an attention module to extract hidden features for the input of the following layer. Then, we use the first attention layer because it focuses on the most relevant information and filters out irrelevant information in a pair of images and questions, and the later layers repeatedly elaborate on the hidden features of the image and question. All codes are implemented using Pytorch 0.4.1 and are available in public.

Results of Task 3: Out-of-distribution Image and Question

The experimental results of task 3 (out-of-distribution image and question) is in Table 7. All combinations of out-of-distribution images (MNIST, SVHN, FashionMNIST, CIFAR10, and TinyImageNet) and questions (20 Newsgroup, Reuter52, and IMDB) are considered in task 3. This situation when both image and question are from out-of-distribution is an extreme case, but we contain task 3 to evaluate the robustness of VQA models on the extreme case that is expected to be detected more easily.

The results show that the attention-based approach and our regularization outperform compared methods. For all VQA models and datasets, attention-based approach shows almost perfect detection of out-of-distribution samples. When we consider that task 3 is rather extreme and easy than task 1 and 2, the superior results are acceptable.

Comparison with Minimization of Attention Variance

Another alternative to make an attention weight uniform is minimization of variance of attention distribution. When the variance of attention weights is minimized, the variance is zero and the all attention weights are same.

In Table 8, we find that minimizing the variance of attention networks (RA-VAR) induces more degradation of VQA accuracy than maximization of our regularization $\sum_{i=1}^{K}\sum_{j=1}^{M}\log(1-A_{ij})$ (RA-LOG). In this study, we prefer to use RA-LOG because anomaly detection in VQA is useless if the VQA accuracy does not conserved.

Although the accuracy drop, RA-VAR is also effective to detect various abnormal samples in test time (Table 7). The results imply that our assumption of attention networks appropriates for anomaly detection in VQA regardless of its implementation.

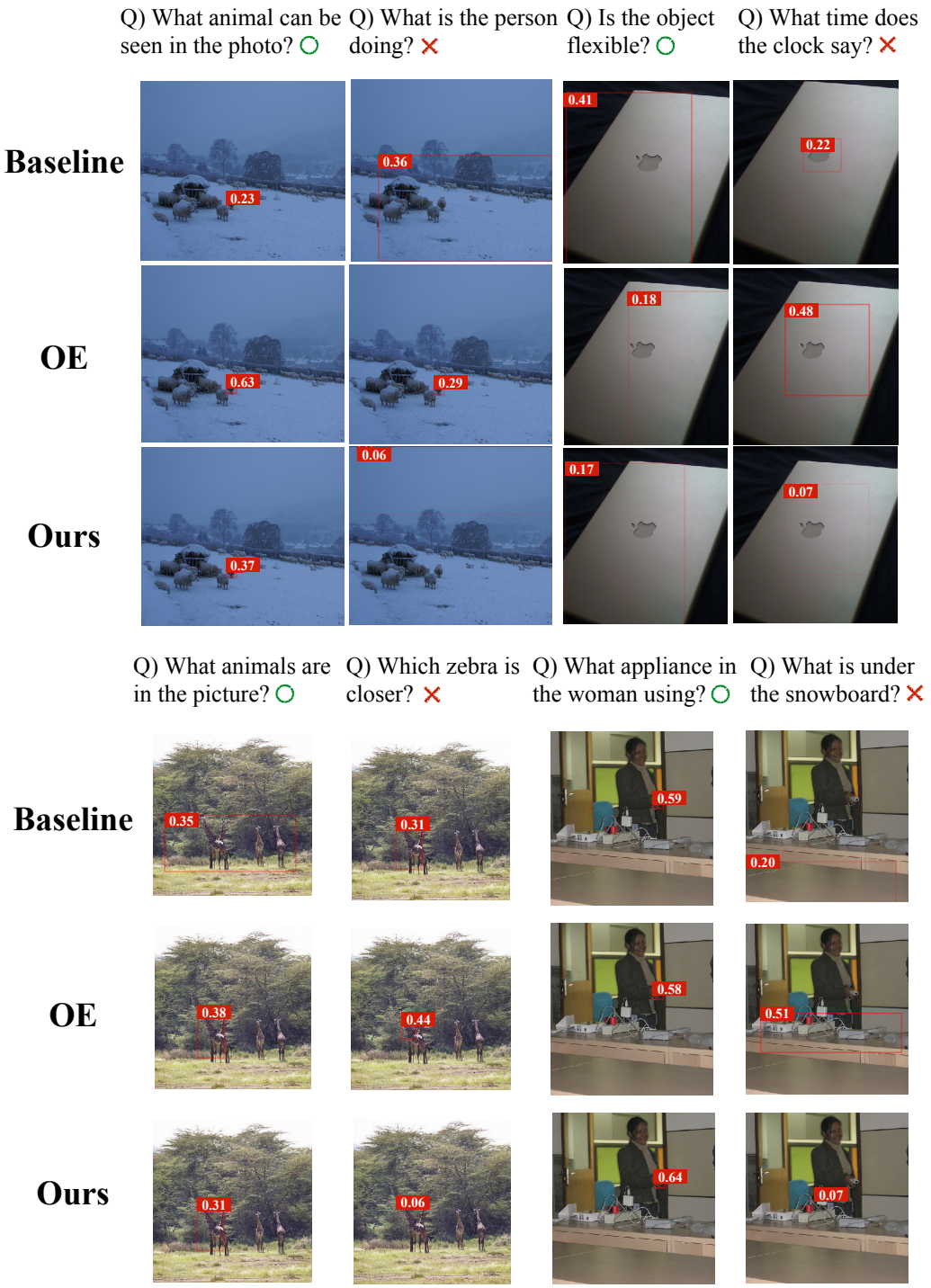
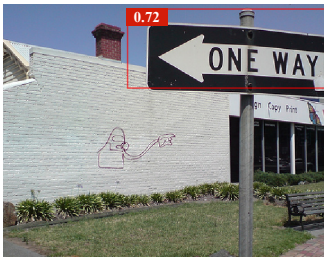


Figure 5: Results of top-1 attention on samples with irrelevant question (1). Baseline (BUTD), baseline + OE, and baseline + our regularization are compared.

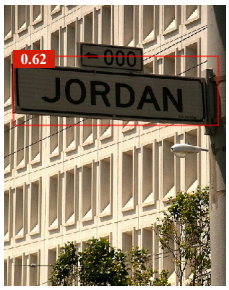
Figure 6: Results of top-1 attention on samples with irrelevant question (2). Baseline (BUTD), baseline + OE, and baseline + our regularization are compared.

Q) What letter and 3 numbers are on the tag?
A) s (15.1 %) **GT)** sv-6260

Q) What is number on the train?
A) 7 (8.2 %) **GT)** 71

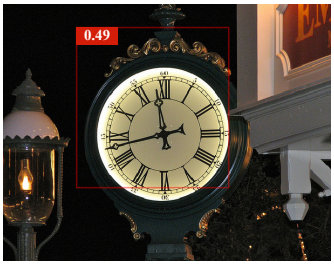


Q) What does the traffic sign say?
A) yield (11.6 %) **GT)** one way



Q) What does the sign say?
A) street (14.4 %) **GT)** jordan

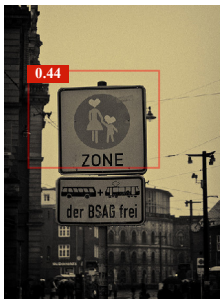
Q) What does the sign read?
A) parking (6.4 %) **GT)** one way



Q) What time does this tell?
A) 7:25 (5.3%) **GT)** 11:43

Q) What does the shirt say?
A) school (14.2 %) **GT)** julie

Q) What is the license plate number?
A) yes (5.7 %) **GT)** p-18368



Q) What does the sign say?
A) stop (16.3 %) **GT)** zone

Q) What does the red sign say?
A) stop (98.2 %) **GT)** detour

Q) What does the sign say?
A) parking (9.9 %) **GT)** wrong way

Q) What time is shown?
A) 1:30 (2.8 %) **GT)** 8:20

Figure 7: Results of top-1 attention on samples with undefined answer (reading texts and numbers). Baseline (BUTD) makes a proper attention, but results in completely different output.

Table 8: Degradation of accuracy by additional training for anomaly detection

Accuracy (%)	Base	OE	RA-VAR	RA-LOG
BUTD	62.55	54.87(-7.68)	61.56(-0.99)	61.91(-0.64)
MHB+ATT	63.33	62.38(-0.95)	62.54(-0.79)	62.77(-0.56)
BAN	63.81	61.92(-1.89)	63.37(-0.44)	63.74(-0.07)

Table 9: Comparison of regularization implementations. RA-VAR: minimization of variance of attention networks. RA-LOG: original implementation in the paper: maximization of $\sum_{i=1}^{K}\sum_{j=1}^{M}\log(1-A_{ij})$

AUROC	BUTD	MHB+ATT	BAN
Image	MAP(T)/RA-VAR/RA-LOG		
MNIST	89.0/97.4/**97.8**	89.9/93.2/**94.7**	99.0/92.2/**100**
SVHN	90.3/97.5/**97.9**	89.7/95.4/**96.2**	100/93.4/**100**
FashionMNIST	89.6/97.4/**97.8**	90.5/94.6/**95.7**	99.9/93.0/**100**
CIFAR10	90.5/97.5/**98.0**	89.9/96.3/**96.9**	100/94.1/**100**
TinyImageNet	92.7/98.5/**99.7**	91.5/99.1/**99.2**	100/96.0/**100**
Question	MAP(T)/RA-VAR/RA-LOG		
20 Newsgroup	78.2/**97.1**/95.5	78.9/**97.4**/92.6	81.7/**93.2**/87.3
Reuters52	76.4/**97.7**/97.0	77.4/**98.6**/94.3	81.7/**93.9**/87.3
IMDB	78.2/**95.1**/92.8	77.9/**95.9**/91.1	78.1/**89.4**/82.5
Irrelevant Q	MAP(T)/RA-VAR/RA-LOG		
VNQ	60.9/88.7/**98.3**	75.8/**99.7**/99.6	73.5/**96.8**/90.0
QRPE	49.6/**89.0**/84.8	56.6/**94.4**/94.0	61.7/**79.5**/63.6

D. Qualitative Analysis of Unanswerable Samples

Examples of Task 4: Irrelevant Question

Experimental results of MAP (T) imply that attention networks in VQA have a bias of highly confident attention to a prominent object regardless of the question relevance. In Fig. 5 and 6, We attach examples of the results of maximum attention scores on samples with irrelevant questions before and after our regularization. Considering the sum of the attention scores is 1.0 over 36 objects (each object has about 0.03 in a uniform distribution), baseline models without regularization or with outlier exposure (Hendrycks, Mazeika, and Dietterich 2019) still give strong attention on a prominent visual object, even though the questions are irrelevant and there is not related visual object in the images. However, after regularization (ours), the attention networks avoid giving strong attention to any prominent object when the question is irrelevant to the image, and the attention scores become close to uniform. Of course, if the question is relevant to the image, our model makes proper and strong attention to the visual object.

Examples of Task 5: Undefined Answer

We further study the samples with undefined answers and attached various examples in Fig. 7. We find that about 40% (1676/4303) of samples with undefined answers require to read texts or numbers in input images. Note that the VQA model makes the right attention on the samples with undefined answers, but makes strange outputs because of their unanswerability by predefined answer candidates. It is a common limitation of recent VQA models, which solve VQA as a prediction problem, such as classification or regression, because there are infinite combinations of text and numbers in the real-world (Singh et al. 2019).

E. Anomaly Detection of Regression-based VQA Models

Although classification-based VQA models show promising results on the VQA v2 challenge (Anderson et al. 2018; Kim, Jun, and Zhang 2018; Yu et al. 2019), regression-based VQA models are also feasible and important approaches in VQA studies. The main differences between classification- and regression-based approaches are that regression-based VQA models take an answer candidate as the input together, compute likelihood $p(\mathbf{a},\mathbf{v},\mathbf{q})$ scores of all answer answer candidates, and select the most likely answer candidate as the prediction:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\mathbb{E}_{p_\mathcal{D}}\left[\log p_\theta\left(\mathbf{a},\mathbf{v},\mathbf{q}\right)\right], \tag{9}$$

$$S(\mathbf{v},\mathbf{q};T) = \max_i p_\theta(\mathbf{a}_i,\mathbf{v},\mathbf{q};T). \tag{10}$$

One can speculate that regression-based VQA models, which jointly consider an image, a question, and the answer, are free of anomaly detection. Thus, we evaluate the robustness of regression-based anomaly detection methods. We modify the

Table 10: Anomaly Detection Performances (task 1, 2, and 4) of regression-based VQA model (NTN).

AUROC	BUTD + NTN
Image	MSP/MSP(T)/OE-MSP(T)/MAP(T)/RA-MAP(T)
MNIST	67.5/84.3/64.3/**93.8**/92.3
SVHN	67.7/85.5/66.0/**94.5**/92.9
FashionMNIST	67.6/85.1/66.3/**94.3**/92.5
CIFAR10	67.8/85.9/66.9/**94.6**/93.1
Tiny	67.6/86.7/74.9/**95.8**/94.1
Question	MSP/MSP(T)/OE-MSP(T)/MAP(T)/RA-MAP(T)
20 Newsgroup	68.9/83.4/70.3/78.1/**92.7**
Reuters52	69.9/83.5/68.4/77.4/**94.0**
IMDB	63.3/80.6/66.9/77.5/**89.1**
Irrelevant Q	MSP/MSP(T)/OE-MSP(T)/MAP(T)/RA-MAP(T)
VNQ	83.8/92.0/76.4/69.9/**96.0** /
QRPE	67.3/70.8/79.8/45.4/**89.1**

model architecture of BUTD based on common approaches, neural tensor networks (NTN) (Bai et al. 2018). Note that although regression-based VQA models compute a likelihood on every answer candidate, cross-modal attention networks are located before the multi-modal feature fusion, and our attention-based anomaly detection can still be used (see the details in (Bai et al. 2018)). The results in Table 10 imply that the regression-based approach still suffers from various anomalies, and attention-based anomaly detection makes the VQA model robust to them.