Adversarial Robustness of Bottleneck Injected Deep Neural Networks for Task-Oriented Communication

Alireza Furutanpey, Pantelis A. Frangoudis, Patrik Szabo, Schahram Dustdar
TU Wien
Distributed Systems Group

***Abstract*—This paper investigates the adversarial robustness of Deep Neural Networks (DNNs) using Information Bottleneck (IB) objectives for task-oriented communication systems. We empirically demonstrate that while IB-based approaches provide baseline resilience against attacks targeting downstream tasks, the reliance on generative models for task-oriented communication introduces new vulnerabilities. Through extensive experiments on several datasets, we analyze how bottleneck depth and task complexity influence adversarial robustness. Our key findings show that Shallow Variational Bottleneck Injection (SVBI) provides less adversarial robustness compared to Deep Variational Information Bottleneck (DVIB) approaches, with the gap widening for more complex tasks. Additionally, we reveal that IB-based objectives exhibit stronger robustness against attacks focusing on salient pixels with high intensity compared to those perturbing many pixels with lower intensity. Lastly, we demonstrate that task-oriented communication systems that rely on generative models to extract and recover salient information have an increased attack surface. The results highlight important security considerations for next-generation communication systems that leverage neural networks for goal-oriented compression.**

***Index Terms*—Task-Oriented Communication, Goal-Oriented Compression, Adversarial Machine Learning, Information Bottleneck**

I. Introduction

Intelligent tasks refer to programs that classical control structures cannot compute tractably or with sufficient precisions, which is common in visual applications, such as image recognition. Deep Learning (DL) has repeatedly demonstrated that it can solve recognition tasks reliably. Unsurprisingly, applications with stringent performance criteria (e.g., remote sensing, video analytics) increasingly offload requests to a remotely deployed large Deep Neural Network (DNN). The pervasiveness of DNNs exposes significant vulnerabilities to adversarial attacks [1]. Another limitation is that continuous offloading of high-dimensional visual data must compete for limited bandwidth, which may lead to network congestion. Task-oriented communication [2] has emerged as a paradigm to meet the need for solving intelligent tasks. Compression for task-oriented communication uses a semantic rate-distortion objective to transmit only the most salient bits. Among the earliest examples is the information bottleneck [3] (IB), which is still the foundation of modern approaches. Notably, IB-based objectives improve adversarial robustness for DNN predictors [4]. The idea is that perturbations are intrinsically redundant information, and the IB objective naturally enhances the robustness of DNNs by learning to discard redundancy more aggressively along their information processing path [5]. However, the established consensus on the value of IB-based objectives is based on the assumption that the networks are *deep*. Yet, task-oriented communication is feasible only when paired with lightweight compression such that the codec computational overhead is offset by the reduced transmission costs [6], [7]. Moreover, there are additional constraints on the encoder design. While a neural encoder may still be wide enough to leverage parallelization from onboard AI accelerators, meeting stringent latency requirements demands reducing the number of sequential operations. Hence, envisioned future communication networks that rely on neural encoding schemes will realistically converge towards *shallow* networks.

979-8-3503-5614-4/24/$31.00 © 2024 IEEE

To this end, this work investigates the robustness of methods that apply an IB-based objective intended for task-oriented communication. Specifically, we apply several common adversarial attacks on recent approaches based on *Shallow Variational Bottleneck Injection* (SVBI) [8]–[12]). SVBI focuses on information necessary only for practically relevant tasks by targeting the shallow representation of foundational models as a reconstruction target in the rate-distortion objective. Our results show that deep networks trained with a traditional IB objective exhibit higher adversarial robustness than SVBI. However, a shallow variational encoder still provides a defense mechanism that results in considerably more robust models than purely discriminative models trained with non-IB objectives. We finalize our experiments by accentuating the increased attack surface of systems that rely on generative models for communicating salient information with a simple attack specifically targeting generative models. In other words, the overall system is more vulnerable even if task-oriented communication is intrinsically more robust than passing messages through conventional channels for downstream tasks. We summarize our contributions as follows:

- Empirically demonstrating that task-oriented communication systems have an increased attack surface.
- Showing that adversarial robustness for task-oriented communication systems requires a study distinct from general research on security for DNNs.
- Determining the role of bottleneck depth for adversarial robustness with IB-based objectives.

We hope our results and insights can facilitate research in securing next-generation communication systems that rely on

otherwise easily exploitable neural networks.

II. Background & Related Work

A. Adversarial Attacks on DNNs

Adversarial attacks represent a significant challenge for deploying AI systems. The susceptibility of DNNs to adversarial examples was first investigated by Szegedy et al. [13], who demonstrated that small, imperceptible perturbations to input data can lead to significant misclassifications. Adversarial attacks are classified into white-box and black-box attacks. White-box attacks assume complete model knowledge, including architecture and gradient calculation, allowing for computing highly effective adversarial samples. In contrast, black-box attacks assume no access to model details and are generally more challenging but more realistic for real-world scenarios. The following briefly describes the attacks we have chosen due to their influence and being subject to numerous follow-up studies. The focus is on white-box attacks due to the open nature of ML research and the popularity of readily available open-source weights for a wide range of tasks.

1) Fast Gradient Sign Method (FGSM): FGSM by Goodfellow et al. [14] efficiently generates adversarial examples by leveraging the gradient of the loss function. FGSM adjusts the input along the gradient's direction, with the perturbation defined as:

$$x_{adv} = x + \epsilon \cdot \operatorname{sign}\left(\nabla_x J(x, y)\right) \tag{1}$$

where x is the input, ϵ controls perturbation magnitude, and $\nabla_x J(x, y)$ is the gradient of the loss concerning the input.

2) Carlini and Wagner (C & W): The attack by Carlini and Wagner [15] minimizes the L_2, L_0, or L_∞ distance between the original input and the adversarial example, and a term that penalizes classifications other than the desired target class using the objective function:

$$J(x') = \alpha \cdot ||x - x'||_p + \beta \cdot \mathcal{L}_{\text{mcls}}(f(x'), y_t) \tag{2}$$

where x' is the perturbed input, α, β balance the terms, and $\mathcal{L}_{\text{mcls}}$ is the missclassification loss. Notably, this attack is shown to be highly effective against networks pre-trained on ImageNet, which are commonly used to finetune by practical recognition tasks.

3) Elastic-Net Attacks on DNNs (EAD): This method [16] is particularly useful for producing sparse perturbations, which can trick DNNs while maintaining minimal changes to the input. It generates adversarial samples by minimizing the objective

$$c \cdot f(\mathbf{x}, t) + \beta \left\|\mathbf{x} - \mathbf{x}_0\right\|_1 + \left\|\mathbf{x} - \mathbf{x}_0\right\|_2^2 \tag{3}$$

where $f(x, t)$ is a target loss function and $c, \beta \geq 0$ are the regularization parameters. EAD's dual-norm optimization is an interesting alternative benchmark for evaluating how variational bottleneck injection handles diverse attack strategies.

4) Jacobian-based Saliency Map Attack (JSMA): The JSMA attack by Papernot et al. [17] constructs adversarial examples by identifying and perturbing input features most critical to the classifier's decision-making process. Unlike gradient-based methods, JSMA uses forward derivatives to create a saliency map, guiding perturbations to specific input features. Given that variational bottleneck techniques may alter feature representations, testing JSMA will allow us to explore how bottleneck injection influences feature saliency and adversarial resilience.

5) Targeting Generative Models: Lastly, we include the attack introduced by Tabacof et al. [18] to demonstrate the increased attack surface of communication systems that deploy generative models. This attack disrupts reconstruction and induce the encoder to produce a completely different target image. This would undermine the potential defensive role of autoencoders in de-noising classifier inputs. Note that the efficacy of the attack towards the autoencoder is irrespective of whether we map the latent to an approximation of the original image (i.e., reconstruction) or use it for some image recognition downstream task [9].

B. Information Bottleneck in Task-Oriented Compression

Using Shannon's rate-distortion (r-d) theory [19], we seek a mapping bound by a distortion constraint from a random variable (r.v.) X to a r.v. Y, minimizing the bitrate of the outcomes of X. More formally, given a distortion measure $\mathcal{D}$ and a distortion constraint D_c, the minimal bitrate is characterized by the *rate-distortion function*:

$$\min_{P_{Y|X}} I(X;Y) \text{ s.t. } \mathcal{D}(X,Y) \leq D_c \tag{4}$$

where $I(X;Y)$ is the mutual information and is defined as

$$\mathrm{I}(X;Y) = \int\!\!\int p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy \tag{5}$$

In lossy image compression, Y is typically an approximate reconstruction of X. This objective lends itself to the Information Bottleneck that maps X to a hidden representation Z, which is minimally informative of X but is also maximally informative about a target prediction task Y. In other words, it is essentially a flavor of the lossy-source coding problem using a different loss as a fidelity measure for the distortion constraint.

1) Deep Variational Information Bottleneck (DVIB): Given ground-truth labels Y from a joint distribution $P_{X,Y}$, the Deep Variational Information Bottleneck objective is to maximize

$$I(Z;Y) - \beta I(Z;X) \tag{6}$$

where β is a Lagrange multiplier. To approximate $I(Z;Y)$ we can apply the conditional cross entropy (CE) $H(P_Y, P_{\tilde{Y}|Z})$. The first term is commonly referred to as the *relevance* and the second as the *complexity*. While the original work [4] considers the complexity term a regularizer, Singh et al. [20] apply it as a rate term to end-to-end train a neural compression model. Dubois et al. [21] generalize the information bottleneck

objective for compression that preserves salient pixels for a set of tasks that share common properties. However, both works rely on deep networks and place the bottleneck at the penultimate layer or require a large pre-trained encoder.

2) Shallow Variational Bottleneck Injection (SVBI): Instead of targeting a particular task Y, SVBI considers a foundational model $\mathcal{M}$, that supports a set of unknown tasks $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_t\}$. Moreover, it partitions $\mathcal{M}$ into two disjoint sets of shallow and deep layers $\mathcal{M} = (\mathcal{T}, \mathcal{H})$, such that for observations X, $\mathcal{M}(X) = \mathcal{T}(\mathcal{H}(X))$, and $\mathcal{H} = H$ is a *shallow* hidden representation of X. Further, assume a codec $\mathrm{c} = (\mathrm{enc}, \mathrm{dec})$, where $\mathrm{dec}(\mathrm{enc}(X)) = \tilde{H}$ is an approximation of H. The idea is that if $\tilde{H}$ is a sufficient approximation of H, then the compressed representation $\mathrm{enc}(X)$ is informative enough of the entire set of tasks $\mathcal{Y}$. While SVBI still uses task performance as a fidelity measure, the compression model is end-to-end optimized using Head Distillation (HD) [22], [23] as the distortion term in the loss function. Figure 1 visually explains the loss function. We refer to our earlier work [8],

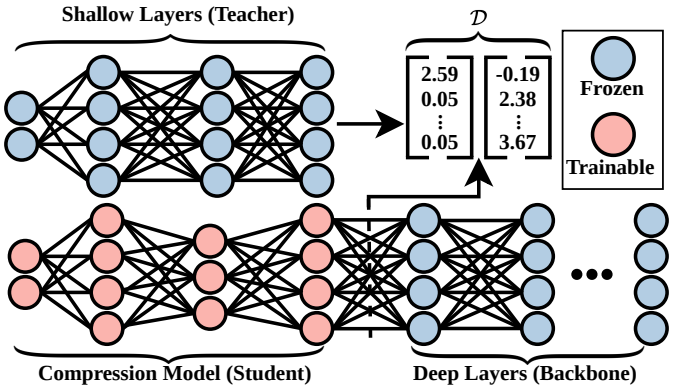

Fig. 1. Head Distillation Loss: The shallow features of pre-trained large models are cut and used as a teacher network to train the compression model.

[9] for a detailed explanation. For this work (i.e., determining adversarial robustness respective bottleneck location depth), it is only relevant that the encoder is shallow with roughly $150,000$ parameters and that the method can significantly reduce bitrate while ensuring task integrity without relying on a labeled dataset. A general downside of lightweight encoders and transmitting information intended to generalize to a broader range of tasks is an increased bitrate relative to deep IB methods.

III. Problem Statement & Methodology

A. Information Bottleneck for Adversarial Robustness

Based on the following two observations, we argue that SVBI should still provide a certain level of adversarial robustness but significantly less than DVIB.

First, the depth, i.e., *the large number of stacked layers* up until the bottleneck, may be an essential reason for the efficacy of adversarial robustness using IB-based objectives. Consider an n-layered feed-forward neural network as a Markov chain of successive representations R_i, R_{i+1} [24]:

$$I(X;Y) \geq I(R_1;Y) \geq \ldots \geq I(R_n;Y) \geq I(\tilde{Y};Y) \quad (7)$$

That is, discerning salient from redundant information is part of transforming an input for prediction. A longer sequence of operations permits the network to process the input with more diverse views. Therefore, we reason that deeper models may benefit more from the IB objective, as they can learn more varied representations for filtering redundant information (i.e., adversarial noise).

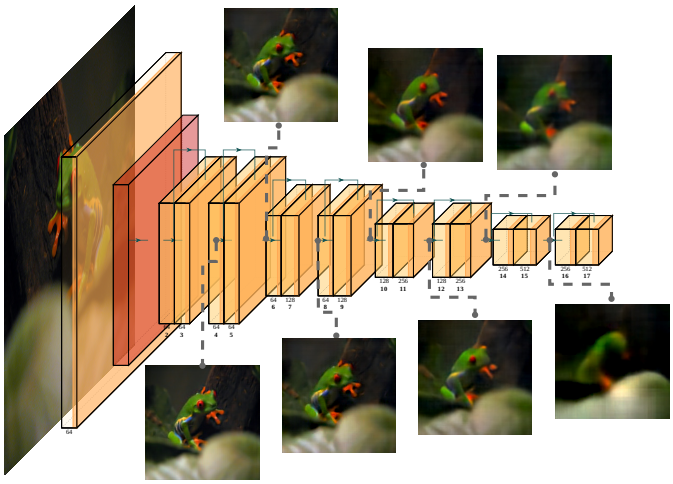

Fig. 2. Filtering Redundant Information for the ImageNet classification task as the network transforms features. For each layer, we trained a separate reconstruction network.

Second is the *task specifity* of the objective. Consider a visual illustration of the information path equation in Figure 2. The frog subset of ImageNet distinguishes between *Tree Frogs*, *Bullfrogs*, and *Tailed Frogs*. Since these frog species have distinct figures and dominant colors, the more delicate characteristics of a tree frog are redundant for ImageNet classification. In SVBI, we place the bottleneck in the first or second marker region, whereas in DVIB, we place it around the last marker. Clearly, when the target task is specifically ImageNet, there is still a considerable amount of redundancy. In other words, when we use an IB-based objective that aims to generalize to a range of tasks, there is more ambiguity to exploit.

B. Attack Surface of Task-Oriented Communication Systems

Figure 3 illustrates a simplified task-oriented communication system that relies on some form of generative method for compression that can extract and recover salient information. Before passing the input to a discriminative prediction model, we process it with a generative compression model. We argue that even if training the goal-oriented neural codec with an IB-based objective improves adversarial robustness against attacks intended for discriminative tasks, we are still increasing the attack surface of our overal communication system due to the generative components. Therefore, even if the system uses

the generative component only for extracting and recovering salient information, exploiting generative components should still be possible, such that it compromises the entire system.

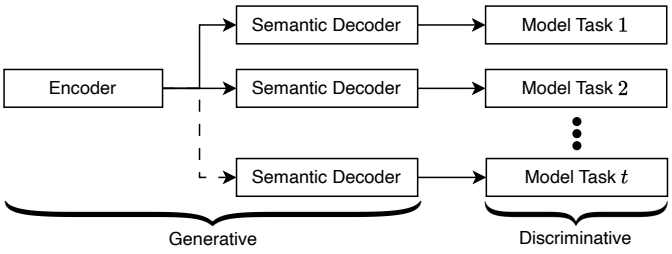

Fig. 3. A simplified overview on a Task-Oriented Communication System. Envisioned systems rely on generative models to encode, transmit and decode salient information for downstream tasks using discriminative models.

C. Adversarial Attacks and Image Perturbations

We generate adversarial samples using the torchattacks [25] library. Except for the Tabacof attack, we create samples for *CIFAR-10*, *SVHN*, and *ImageNet64* (i.e., downsampled ImageNet but still using all original 1000 classes). Notably, we choose JSMA as it may provide a different perspective on model vulnerability by perturbing specific input features. However, JSMA has high memory requirements, which we cannot accommodate with our limited resources for ImageNet64. Therefore, we implement a modified version of JSMA (JSMAOnePixel) that is inspired by [26]. The OnePixel variant identifies only a single pixel with the highest impact on each iteration. Still, as Figure 4 exemplifies, the final perturbed image is comparable between JSMA and JSMAOnePixel.

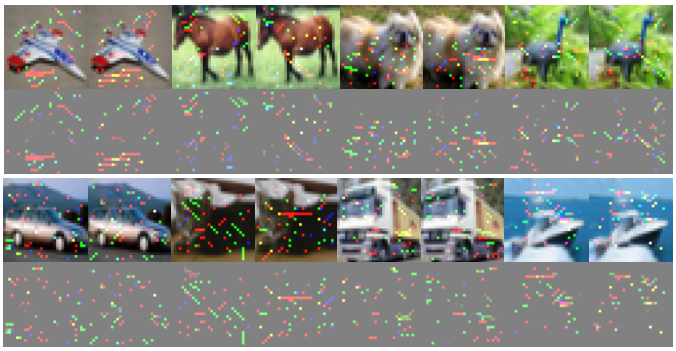

Fig. 4. In pairs, comparing JSMA (left) with JSMAOnePixel variant (right).

IV. Evaluation

A. Adversarial Attacks and Image Perturbations

The aim is to design experiments that yield adequate empirical evidence to conclude the baseline robustness we may expect for the types of compression models used for task-oriented communication. We perform the attacks described in Section II-A for each model and task separately using the datasets described in Section III-C. This is with the exception of the Tabacof attack, where we use MNIST for simplicty as the purpose is to demonstrate the widened attack surface incurred by task-oriented communication.

B. Training Models with (Shallow) Bottlenecks

We train three sets of models, i.e., baseline models with standard log-loss, models with a shallow bottleneck (SVBI), and models with a deep bottleneck (DVIB). We perform DVIB and SVBI as described in [20] and [8], respectively. For DVIB, we place a bottleneck at the penultimate layer and use a log-loss for the distortion term in the objective function. For SVBI, we follow the "blueprint" encoder design that replaces the layers until the first high-level block of the network (roughly 1% of the total model parameters) with a small variational autoencoder. We experimentally determine the lowest possible bitrate for both bottleneck approaches without sacrificing prediction performance.

Table I summarizes the model performances. The bits per pixel (bpp) is a lower bound we have empirically determined for a bottleneck injected model to perform (near-)lossless prediction as defined in [8], [9]. Naturally, DVIB has much lower bitrates for reasons described in Section III-A.

TABLE I
Models Prediction and Compression Performance

Dataset	Acc@1 [%]	Bpp (SVBI)	Bpp (DVIB)
MNIST	97.36 ± 1.77	0.0829	0.0161
CIFAR-10	85.25 ± 1.40	0.5677	0.0308
SVHN	94.04 ± 0.69	0.4321	0.0086
ImageNet64	49.36 ± 1.12	1.2673	0.0115

C. Comparing Bottleneck Placements

Table II summarizes the effect on the adversarial samples represented as percentage points (lower is better). Unsurprisingly, base models trained using a standard log-loss have a significant drop in accuracy. Relative to the accuracy on the unperturbed dataset (Table I), all attacks completely tank the model performance. In particular, for the SVHN task the performance is at times worse than random guessing. As conjectured in Section III-A, SVBI generally provides less adversarial robustness than DVIB across all datasets. Notably, task complexity apparently influences the gap in adversarial robustness between SVBI and DVIB. Still, SVBI exhibits considerably higher adversarial robustness over the baseline. Additionally, notice that the model depth on the base model does not considerably affect adversarial robustness. However, for the DVIB model, depth seems to correlate positively with adversarial robustness. Presumably, since deeper models have longer information paths, end-to-end training models with an IB objective have more opportunities to discard information that does not contribute to task performance gradually.

D. Analyzing Pixel Perturbations

We observe that IB-based objectives exhibit stronger robustness against attacks that focus on a small subset of salient pixels with strong intensity than attacks that perturb many pixels with smaller intensity. Moreover, similarly to the original work on deep variational IB [4], we observe that attacks targeting IB-based models perturb pixels considerably

TABLE II
COMPARING PREDICTION PERFORMANCE DECREASE (% POINTS; LOWER IS BETTER) BETWEEN OBJECTIVES.

		CIFAR-10			SVHN			ImageNet64		
Model	Attack	Base	SVBI	DVIB	Base	SVBI	DVIB	Base	SVBI	DVIB
ResNet-18	FGSM	74.5621	48.7298	39.513	69.9521	55.8298	48.5728	37.7602	31.8935	28.9807
	EAD	85.5256	9.6447	8.8592	89.9427	19.8432	13.5824	35.4344	9.2381	8.0993
	C&W	87.0232	7.7732	6.7682	92.2149	22.9992	18.3259	37.9903	12.5742	10.2117
	JSMA	87.6210	20.7807	17.5784	91.5810	14.2348	11.1283	36.3821	11.1868	10.1935
ResNet-50	FGSM	68.5621	42.8942	34.0803	68.2679	53.2118	45.1977	39.6985	28.9273	23.1021
	EAD	85.1400	9.1258	7.8592	89.3852	18.0232	11.4375	32.6361	7.0377	4.8375
	C&W	88.9231	7.2009	6.5408	93.3284	18.0931	15.3259	34.1083	9.9654	5.4281
	JSMA	85.1010	19.6333	16.0549	90.2838	13.9125	10.1283	34.4847	10.1351	5.5213
ResNet-101	FGSM	69.3189	40.8912	32.1534	66.2082	40.4817	42.9004	38.4451	24.0620	20.9997
	EAD	87.6557	8.0322	7.8592	88.3294	16.4385	9.5729	32.4148	6.1124	3.0489
	C&W	87.4633	7.1819	6.0018	92.1923	17.3284	12.2482	38.0200	8.7985	2.9663
	JSMA	86.8781	19.439	15.2608	94.2933	11.2814	7.9833	36.6825	10.0382	1.4762

more than non-IB-based models. Nevertheless, since relative values align across all models (i.e., attacks behave comparably regardless of the model depth or objective), the following reports average values due to space constraints.

Figure 5 visualizes the L_0 norm by attack averaged over test sets, i.e., it measures how many pixels an attack has perturbed. While FGSM perturbs nearly all pixels, JSMA only perturbs

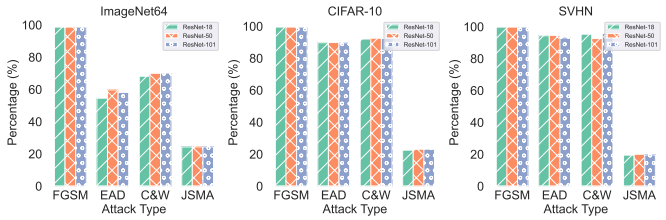

Fig. 5. Average Percentage of pixels perturbed by an adversarial attack. More complex tasks tend to have more salient pixels.

roughly 20% of the pixels. More interestingly, EAD and C&W perturb fewer pixels for ImageNet than for the simpler tasks. Generally, more complex tasks with many labels rely on more fine-grained information, where just a small subset of salient pixels can influence the decision boundaries. Figure 6 summarizes the magnitude of perturbations. Notice that JSMA

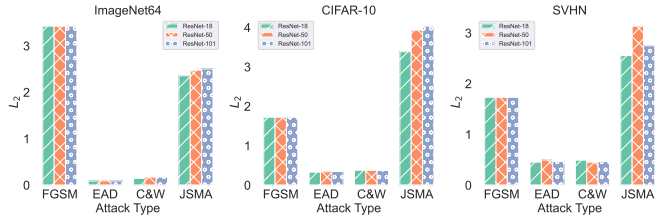

Fig. 6. The average L_2 measures the magnitude of perturbations. FSGSM and JSMA incur considerably higher perturbation than EAD and C&W.

has a larger total magnitude in total perturbation than FGSM despite JSMA focusing on a smaller subset of pixels. The reason becomes apparent when examining the L_∞ norm in Figure 7. JSMA is more "pixel-efficient" by focusing on the

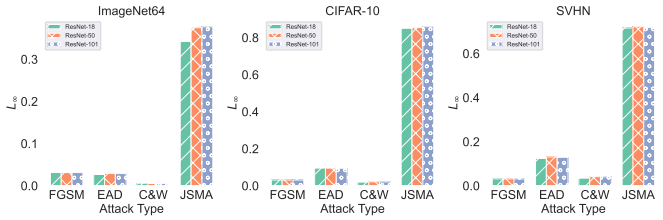

Fig. 7. Average L_∞ measure to quantify the magnitude of perturbation. JSMA perturbs a small number of pixels with high intensity.

most salient pixels but relies on high-magnitude perturbations. Figure 8 visually compares JSMA and FGSM. While FGSM perturbs a large number of pixels, they are only faintly visible. Conversely, JSMA has clearly visible perturbations. This

Fig. 8. Comparing magnitudes of pixels between JSMA and FGSM.

observation is consistent with the general objective of goal-oriented communication, which is to focus on the most salient information. Therefore, it may be reasonable to emphasize evaluating defense strategies for task-oriented communication against less perceptible attacks.

E. Targeted Autoencoder Attack

As described in Section III-B task-oriented communication networks are powered by generative models for communication and discriminative models for high-level downstream

tasks, which incrases the attack surface. We show this by performing the Tabacof [18] attack described in Section II-A5 and summarize the results in Table III. We choose the label

TABLE III
TABACOF ATTACK

	Base		DVIB	
Model	Acc@1	#Hits	Acc@1	#Hits
Resnet-18	61.6	927	52.26	1802
Resnet-50	76.57	879	72.66	1126
Resnet-101	33.17	448	34.73	1641

"1" as the target, and the *hits* column indicates how often the model has predicted "1" after the attack. Since the models have near-perfect accuracy on MNIST, and the test set has 10 000 samples that are uniformly distributed, we can infer the efficacy of the attack by the increase of predictions of "1". Figure 9 shows an example with a curated sample of perturbed images. ResNet-101 is noticeably less robust toward

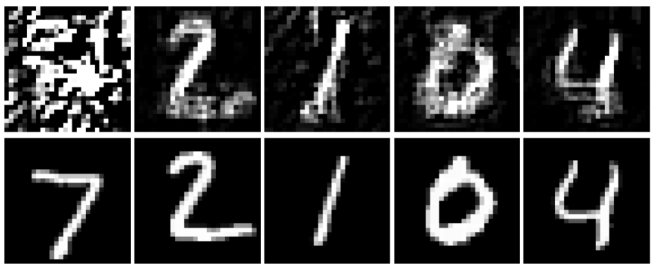

Fig. 9. Base images (bottom) and corresponding perturbation using the Tabacof attack against DVIB (top). The target label is "1". Examples depicting the numbers two and four contain the most clearly visible perturbations to match this target. The leftmost example showcases a "failed" attack, where the network will likely missclassify the input, but not hit the intended target.

the attack. We explain the discrepancy by the simplicity of the task and dataset size. Since ResNet-101 is significantly larger, the model may have been fitted to the samples, making it more susceptible to even slight perturbations. Still, when comparing the performance of ResNet-18 and ResNet-50 shows that the DVIB model is considerably *less robust* than the baseline model. Notably, all DVIB models have a substantial increase in predicting the target label, indicative of the attack's efficacy.

V. CONCLUSION

This work investigated the role of IB-based objectives for task-oriented communication systems and their implications for adversarial robustness. We have shown that such approaches provide a degree of resilience against attacks targeting downstream tasks. However, the reliance on generative models for extracting and recovering salient bits introduces a new attack surface. An attacker may bypass security for prediction models by targeting the semantic layer of a communication system. While research in adversarial attacks that views generative and discriminative models in isolation is essential, we find it indispensable for research to examines thei interaction holistically. Additionally, an promising direction is in methods that quantify the trade-off between generalization and its adversarial robustness as an objective function for end-to-end optimization of goal-oriented codecs.

ACKNOWLEDGMENT

We thank Alexander Knoll for providing us with the hardware infrastructure and Valentin Flunkert for his support.

REFERENCES

[1] N. Akhtar and A. S. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018.
[2] D. Gündüz *et al.*, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, 2023.
[3] N. Tishby *et al.*, "The information bottleneck method," *CoRR*, vol. physics/0004057, 2000.
[4] A. A. Alemi *et al.*, "Deep variational information bottleneck," *CoRR*, vol. abs/1612.00410, 2016.
[5] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *CoRR*, vol. abs/1703.00810, 2017.
[6] A. Mostaani *et al.*, "Task-oriented communication design at scale," *IEEE Transactions on Communications*, 2024, preprint.
[7] J. Peng *et al.*, "Task-oriented multi-user semantic communication with lightweight semantic encoder and fast training for resource-constrained terminal devices," *IEEE Wireless Communications Letters*, vol. 13, no. 9, pp. 2427–2431, 2024.
[8] A. Furutanpey *et al.*, "Frankensplit: Efficient neural feature compression with shallow variational bottleneck injection for mobile edge computing," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 10 770–10 786, 2024.
[9] ——, "Fool: Addressing the downlink bottleneck in satellite computing with neural feature compression," *CoRR*, vol. abs/2403.16677, 2024.
[10] Y. Matsubara *et al.*, "Supervised compression for resource-constrained edge computing systems," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.
[11] Z. Yuan *et al.*, "Split computing with scalable feature compression for visual analytics on the edge," *IEEE Transactions on Multimedia*, vol. 26, pp. 10 121–10 133, 2024.
[12] I. Harshbarger *et al.*, "Condar: Context-aware distributed dynamic object detection on radar data," in *Proc. IEEE MILCOM*, 2024.
[13] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. ICLR*, 2014.
[14] I. J. Goodfellow *et al.*, "Explaining and harnessing adversarial examples," in *Proc. ICLR*, 2015.
[15] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symposium on Security and Privacy*, 2017.
[16] P.-Y. Chen *et al.*, "EAD: Elastic-net attacks to deep neural networks via adversarial examples," in *Proc. AAAI*, 2018.
[17] N. Papernot *et al.*, "The limitations of deep learning in adversarial settings," in *Proc. IEEE EuroS&P*, 2015.
[18] P. Tabacof *et al.*, "Adversarial images for variational autoencoders," *CoRR*, vol. abs/1612.00155, 2016.
[19] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *IRE National Convention Record*, vol. 4, 1959.
[20] S. Singh *et al.*, "End-to-end learning of compressible features," *CoRR*, vol. abs/2007.11797, 2020.
[21] Y. Dubois *et al.*, "Lossy compression for lossless prediction," in *Proc. NeurIPS*, 2021.
[22] M. Sbai *et al.*, "Cut, distil and encode (CDE): Split cloud-edge deep inference," in *Proc. IEEE SECON*, 2021.
[23] Y. Matsubara *et al.*, "Distilled split deep neural networks for edge-assisted real-time systems," in *Proc. HotEdgeVideo@MobiCom*, 2019.
[24] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. 2015 IEEE Information Theory Workshop (ITW)*, 2015.
[25] H. Kim, "Torchattacks: A pytorch repository for adversarial attacks," *arXiv preprint arXiv:2010.01950*, 2020.
[26] K. Song, "Adversarial.js," https://kennysong.github.io/adversarial.js/, accessed: 2024-07-29.