Robust Deep Learning via Layerwise Tilted Exponentials

Bhagyashree Puranik [1] **Ahmad Beirami** [2] **Yao Qin** [1] **Upamanyu Madhow** [1]

Abstract

State-of-the-art techniques for enhancing robustness of deep networks mostly rely on empirical risk minimization. In this paper, we propose a complementary approach aimed at enhancing the "signal-to-noise ratio" at intermediate network layers, loosely motivated by the classical communication-theoretic model of signaling in a noisy channel. We seek to learn neuronal weights which are "matched" to the layer inputs by supplementing end-to-end costs with a tilted exponential (TEXP) objective function which depends on the activations at the layer outputs. We show that TEXP learning can be interpreted as maximum likelihood estimation of "matched filters" under a Gaussian model for "data noise." TEXP inference is accomplished by replacing batch norm by a tilted softmax enforcing competition across neurons, which can be interpreted as computation of posterior probabilities for the signaling hypotheses represented by each neuron. We show, by experimentation on standard image datasets, that TEXP learning and inference enhances robustness against noise, other common corruptions and mild adversarial perturbations, without requiring data augmentation. Further gains in robustness against this array of distortions can be obtained by appropriately combining TEXP with adversarial training.

1. Introduction

Standard training of deep neural networks through empirical risk minimization is well known to lack robustness against a variety of distortions, including noise, distribution shifts (Hendrycks & Dietterich, 2018; Dodge & Karam, 2017), and adversarial attacks (Szegedy et al., 2014; Goodfellow et al., 2015; Carlini & Wagner, 2017). In order to improve model robustness, one of the most common solutions is to perform data augmentation. For example, adversarial training (Madry et al., 2018), which augments the training data with generated adversarial examples (corresponding to the current realization of the network parameters), is one of the most effective adversarial defenses against adversarial attacks. In addition, different types of data augmentation have also been shown to effectively improve robustness against natural corruptions (Cubuk et al., 2019; Hendrycks et al., 2020; Qin et al., 2023).

In this paper, we propose and explore a strategy for enhancing robustness based on detection and estimation theoretic concepts (motivated by their success in fields such as wireless communication systems), in a manner that is *complementary* to standard empirical risk minimization, with or without data augmentation. In communication theory, the receiver tries to match the incoming signal against a number of possible signal template, each corresponding to a different message. For signaling in Gaussian noise, correlating against these signal templates, often called matched filters, maximizes the signal-to-noise ratio, and the posterior probability of each possible transmitted signal is obtained by feeding suitably scaled matched filter outputs to a softmax layer. Our proposed approach here is to apply these ideas in enhancing "signal-to-noise ratio" at intermediate layers of a neural network, so that the outputs are more resilient to "data noise." Unlike communication systems, we do not have a known set of messages and corresponding transmitted symbols. Rather, we seek to learn neuronal weights at a given layer which are well matched to the set of incoming input patterns, so that for each strong input, a fraction of neurons fire strongly. We accomplish this here by adding layer-wise costs based on tilted exponentials, which we show in Section 3 can be interpreted as maximum likelihood estimation of "matched filter" signal templates under Gaussian noise. For inference, we replace batch norm by a tilted softmax, again motivated by our interpretation of neurons as providing competing signal templates. Our framework allows us to vary the amount of "data noise" we expect during training (smaller if we are training with clean data) and during inference (bigger if we wish to be robust against out of distribution noise). We term a layer designed in this fashion as a tilted exponential (TEXP) layer.

We report here on promising preliminary results for CIFAR-

[1]Department of ECE, University of California Santa Barbara, USA [2]Google Research, USA. Correspondence to: Bhagyashree Puranik <bpuranik@ucsb.edu>.

2nd AdvML Frontiers workshop at 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.
Copyright 2023 by the author(s).

10 (Krizhevsky et al., 2009) obtained by replacing the first layer of a VGG-16 (Simonyan & Zisserman, 2014) network by a TEXP layer. We obtain increased robustness against noise, other common corruptions and mild adversarial perturbations *without requiring data augmentation.* Additional performance gains are obtained by supplementing TEXP adaptation with adversarial training.

2. Related Work

Disparity between the data observed during training and testing phases is a common phenomenon, highlighting the significance of robustness in generalizing to out-of-distribution (OOD) samples. To address this challenge, various methodologies such as in (Schneider et al., 2020; Calian et al., 2021; Kireev et al., 2022) have been proposed for combating common corruptions, with many employing OOD data augmentations (Zhang et al., 2017; Cubuk et al., 2019; Hendrycks et al., 2020). Among the state-of-the-art methods is AugMix by (Hendrycks et al., 2020), which enriches the training images by incorporating a composition of randomly sampled augmentations, to generate a diverse set of augmented images. A consistency loss function supplements the training, which enables smoother network responses. Consistency regularizers have shown to be promising in several other works as well (Tack et al., 2022; Huang et al., 2022).

A complementary set of works demonstrate that adversarial training leads to better robustness against some corruptions. Gilmer et al. (2019) show connections between robustness to adversarial perturbations and distributions shifts, in particular, due to Gaussian noise. Their findings indicate that in order to enhance an alternate concept of adversarial robustness, it is necessary to reduce error rate under high levels of additive noise. Towards making this connection more concrete, Yi et al. (2021) measure shifts between distributions using the Wasserstein distance and analytically prove that an adversarially trained model generalizes well on OOD data. Furthermore, they show that using pre-trained robust models and fine-tuning leads to better generalization on OOD downstream tasks. However, finding techniques that work well for various different kinds of OOD corruptions, particularly without heavy data augmentation, remains challenging. Yin et al. (2019) find that adversarial training and Gaussian noise augmentation improve robustness against certain corruptions like other types of noise and blurs while degrading the performance under *low frequency* corruptions like fog and contrast. They argue that a diverse set of augmentations may be required to combat such trade-offs. Our TEXP method shows promise in achieving broad spectrum robustness without data augmentations.

Our approach of adding layer-wise costs is motivated by recent work (Cekic et al., 2022), which argues that targeting sparse, strong activations at intermediate network layers can increase robustness. Cekic et al. (2022) accomplish this by using Hebbian/anti-Hebbian (HaH) training at the intermediate layers, in which neurons which are more active for an input are promoted towards the input ("fire together, wire together"), while neurons which are less active are demoted away from the input, and by using divisive normalization (enabling smaller outputs to be attenuated by larger outputs) instead of batch norm for inference. In contrast to the neuroscientific motivation in HaH (Cekic et al., 2022), our TEXP training and inference approach is derived from communication-theoretic foundations. While our approach also biases in favor of larger activations, our framework leads to smoother objective functions, and our best schemes substantially outperform the benchmarks in (Cekic et al., 2022).

Prior work on tilted exponentials demonstrates that applying TEXP costs to empirical risk minimization objective function can provide fairness and robustness benefits in a multitude of machine learning problems (Li et al., 2021; 2023). In fact, exponential tilting is well-known in statistics for rejection sampling, rare-event simulation, saddle-point approximation (Butler, 2007), and importance sampling (Siegmund, 1976). It is also at the heart of Chernoff bounds (Dembo & Zeitouni, 2009), as well as analyzing atypical events in information theory (Beirami et al., 2018). Exponential tilting has also appeared as a smoothing method to maximum in optimization literature (Kort & Bertsekas, 1972; Pee & Royset, 2011; Liu & Theodorou, 2019).

To the best of our knowledge, this is the first work to show the benefits of TEXP costs at intermediate layers of a neural network. Unlike prior work on exponential tilting, which is motivated by connections to large deviations, our proposal of layer-wise TEXP costs is motivated by maximum likelihood estimation of signal templates.

3. Learning Signal Templates via TEXP

We provide here a communication-theoretic motivation for training and inference in a TEXP layer, and then describe how to incorporate these insights into a neural network architecture in the next section.

A classical model in communication theory is to model the received signal as one of M possible transmitted signals, corrupted by white Gaussian noise. Our TEXP model arises from fitting this model to the input $\mathbf{x}$ to a neural model.

Modeling $\mathbf{x}$ as the observation in an M-ary hypothesis testing problem, under hypothesis $\{H_i\}_{i\in[M]}$, ($[M] := \{1, \ldots, M\}$), we have

$$H_i : \mathbf{x} = \mathbf{s}_i + \mathbf{n} \tag{1}$$

where $\{\mathbf{s}_i\}_{i\in[M]}$ are the possible signals, and $\mathbf{n}$ is white

Gaussian noise with variance σ^2 per dimension. While such a model is not expected to be accurate for the layer input in a neural model, fitting it to data provides an approach for learning neural weights such that, for each input, it is likely that there is a subset of neurons well matched to it. The parameter σ^2 may be viewed as "data noise," acknowledging that the input $\mathbf{x}$ may not fit any of the templates we learn.

TEXP Training. We wish to learn the signal templates $\theta = \{\mathbf{s}_i\}_{i\in[M]}$ from data, for a given TEXP layer. The likelihood function conditioned on θ and H_i is given by

$$L_\theta(\mathbf{x}|H_i) = \exp\left(\frac{1}{\sigma^2}(\langle \mathbf{x}, \mathbf{s}_i\rangle - ||\mathbf{s}_i||^2/2)\right), \quad (2)$$

for $i \in [M]$. This likelihood function is the Radon-Nikodym derivative of the conditional distribution of H_i with respect to that of a "noise only" dummy hypothesis $\mathbf{x} = \mathbf{n}$. Assuming that all signal templates have equal energy, we can drop the $||\mathbf{s}_i||^2/2$ terms from (2) to obtain the simplified expression, for all $i \in [M]$:

$$L_\theta(\mathbf{x}|H_i) = \exp\left(\frac{1}{\sigma^2}\langle \mathbf{x}, \mathbf{s}_i\rangle\right). \quad (3)$$

Averaging over these conditional likelihoods (3), the likelihood of $\mathbf{x}$ is now obtained as a sum of tilted exponentials:

$$L_\theta(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M}\exp\left(\frac{1}{\sigma^2}\langle \mathbf{x}, \mathbf{s}_i\rangle\right) = \frac{1}{M}\sum_{i=1}^{M}\exp\left(ta_i\right), \quad (4)$$

where $t = \frac{1}{\sigma^2} > 0$ is the tilt parameter and $a_i = \langle \mathbf{x}, \mathbf{s}_i\rangle$ is the activation for the ith neuron. The corresponding log likelihood is the tilted exponential objective function:

$$T_\theta(\mathbf{x}) = \log L_\theta(\mathbf{x}) = \log \frac{1}{M}\sum_{i=1}^{M}\exp\left(ta_i\right). \quad (5)$$

Maximization of the objective function (5), added across training data points, over θ provides a maximum likelihood estimate of the signal templates.

The gradient of this objective function is given by

$$\nabla_\theta T_\theta = t\sum_{i=1}^{M}\frac{e^{ta_i}}{\sum_{j=1}^{M}e^{ta_j}}\nabla_\theta a_i = t\sum_{i=1}^{M}\text{Softmax}_i(t\mathbf{a})\nabla_\theta a_i, \quad (6)$$

where $\mathbf{a} = \{a_i\}_{i\in[M]}$ and $\text{Softmax}_i(\cdot)$ is the ith index of the softmax output. Since larger activations are weighted more via the tilted softmax, gradient ascent corresponds to increasing larger activations further: since the signal templates are normalized, this requires aligning the templates yielding larger activations more closely with the input.

Additional competition among the signal templates seeking to fit an input can be created by imposing a *balance* constraint in which the mean of the signal templates is set to zero. That is, we replace $\mathbf{s}_i$ by $\mathbf{s}_i - \bar{\mathbf{s}}$, for $i \in [M]$, where $\bar{\mathbf{s}} = (1/M)\sum_{i=1}^{M}\mathbf{s}_i$. Analogous to (5), this corresponds to the balanced tilted exponential objective function

$$T_\theta^{\text{bal}}(\mathbf{x}) = \log \frac{1}{M}\sum_{i=1}^{M}\exp\left(t(a_i - \bar{a})\right),$$

where $\bar{a} = (1/M)\sum_{i=1}^{M}a_i$ is the mean activation of all neurons. The corresponding gradient is given by

$$\nabla_\theta T_\theta^{\text{bal}} = t\sum_{i=1}^{M}\left(\text{Softmax}_i(t\mathbf{a}) - 1/M\right)\nabla_\theta a_i. \quad (7)$$

Now, in addition to trying to make large activations larger, we wish to make small activations (i.e., such that tilted softmax is smaller than $1/M$) smaller.

TEXP Inference. Once we have estimates of the signal templates $\{\mathbf{s}_i\}$, inference based on a data point $\mathbf{x}$ consists of computing the posterior probability of each hypothesis. For hypothesis H_i, this posterior probability is given by

$$p_i(\mathbf{x}) = \frac{L_\theta(\mathbf{x}|H_i)P(H_i)}{\sum_{j=1}^{M}L_\theta(\mathbf{x}|H_j)P(H_j)} = \frac{\exp\left(\frac{1}{\sigma^2}\langle \mathbf{x}, \mathbf{s}_i\rangle\right)}{\sum_{j=1}^{M}\exp\left(\frac{1}{\sigma^2}\langle \mathbf{x}, \mathbf{s}_j\rangle\right)}. \quad (8)$$

Setting $t = \frac{1}{\sigma^2}$, the M-dimensional output corresponding to $\mathbf{x}$ is obtained via the softmax as, for $i \in [M]$:

$$p_i(\mathbf{x}) = \text{Softmax}_i(t\mathbf{a}). \quad (9)$$

The value of σ^2 used during inference using (8) may be different from that for training as in (5). In particular, we may use a smaller value of σ^2 (higher t) during training, where we might be learning from clean data, or from data that we have perturbed in a controlled manner. On the other hand, during inference, we may use a higher value of σ^2 (lower t) in order to accommodate data noise due to a variety of distortions that were not present during training. Note that TEXP inference (8) is unaffected by whether or not the signal templates are balanced, since balancing corresponds to subtracting the same constant from each activation.

4. TEXP as a Neural Network Layer

We now translate these ideas to layer-wise training in a CNN. Our approach is to modify the standard convolution layers of a baseline CNN by replacing conventional ReLU and batch normalization layers by a tilted softmax layer, analogous to computing posterior probabilities in TEXP inference. Each TEXP layer also contributes its own TEXP objective to the training cost, supplementing a standard end-to-end cost.

Similarly to (Cekic et al., 2022), we implicitly normalize the convolution filter weights to unit ℓ_2 norm, in order to enforce fair competition across the signal templates represented by each filter. Given a filter $\mathbf{w}^k$ at a TEXP layer with K

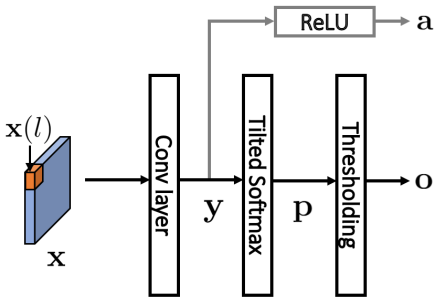

Figure 1. The illustration of a TEXP layer.

convolution filters, and a patch of input $\mathbf{x}(l)$ at the spatial location l, the corresponding output of the k-th filter at location l is computed as a tensor inner product as follows

$$y^k(l) = \frac{\langle \mathbf{x}(l), \mathbf{w}^k \rangle}{||\mathbf{w}^k||_2}. \tag{10}$$

For example, for CIFAR-10 images fed to a VGG-16 model, the first convolution block consists of $K = 64$ filters, each a 3×3 kernel with stride and padding of 1. Thus, we have $L = 32 \times 32 = 1024$ spatial locations and corresponding input patches, where $l \in [L]$ in (10). Let us index the convolution layer outputs, across all filters and spacial locations, by $y_i, i \in [M]$. For this example, we apply TEXP with $M = 32 \times 32 \times 64$.

Post the convolution, we pass the convolution outputs through a Tilted Softmax (TS) to obtain:

$$p_i = \text{Softmax}_i(t_{\text{inf}}\mathbf{y}),$$

where $\mathbf{y} = \{y_i, i = 1, 2, \ldots, M\}$, t_{inf} is the tilt parameter and M is total number of scalar outputs of the TS layer across all filters and spatial locations. We reindex the post softmax outputs p_i by filter k and spatial location l as $p^k(l)$, and use these notations interchangeably.

Further, borrowing another idea from (Cekic et al., 2022), a filter-specific adaptive thresholding is performed to obtain the TEXP layer outputs:

$$o^k(l) = \begin{cases} p^k(l) & \text{if } p^k(l) \geq \tau_k \\ 0 & otherwise \end{cases} \tag{11}$$

where outputs at all locations, arising due to filter k, are subjected to the filter-specific threshold τ_k. The thresholds are set such that for every image, the thresholding block permits only a certain fraction of the activations, while nullifying the rest. For instance, we set τ_k adaptively such that 20% of the outputs are activated for each image, and each filter.

TEXP objective. Each TEXP layer l is associated with its own TEXP objective T_l which is combined with the end-to-end discriminative training cost as follows:

$$J = J_{e2e} - \sum_{l \in \mathcal{T}} \alpha_l T_l, \tag{12}$$

where J_{e2e} is the end-to-end cost function (taken to be a standard discriminative cost in our evaluations here), $\mathcal{T}$ indexes the set of TEXP layers, T_l is the TEXP objective for layer $l \in \mathcal{T}$ and $\alpha_l > 0$ are hyperparameters that determine the relative importance of the TEXP objective compared to the end-to-end cost.

Dropping the subscript l for simplicity of exposition, the TEXP objective for a given layer in $\mathcal{T}$ is given by

$$T = \frac{1}{t} \log \left(\frac{1}{M} \sum_{i=1}^{M} \exp(t a_i) \right) \tag{13}$$

and that for the *balanced* TEXP objective is given by

$$T_{\text{bal}} = \frac{1}{t} \log \left(\frac{1}{M} \sum_{i=1}^{M} \exp(t(a_i - \bar{a})) \right)$$

where $a_i = \text{ReLU}(y_i)$ are the convolution outputs across all filters and spatial locations in the layer l, passed through a ReLU function, M denotes the number of such scalar outputs, t denotes the tilt parameter for the tilted objective, and $\bar{a} = (1/M) \sum_{i=1}^{M} a_i$ denotes the mean of all the post-ReLU activations in the layer. Note that the t_{inf} in the tilted softmax inference is smaller than the tilt t used during training in (13). All TEXP layers could potentially have different tilts in the layer objective and softmax, and layer weights α_l.

5. Experimental Evaluation

The primary focus in our experiments is on the CIFAR-10 standard and corruption datasets with the VGG-16 model as the baseline architecture, where we show significant gains in robustness from tuning a single TEXP layer. We also show gains in robustness from applying TEXP to 6 layers of ResNet; we have not fine-tuned hyperparameters here, and provide these results only to illustrate the applicability of our approach to a variety of architectures and to multiple layers.

Baselines. We obtain two baseline VGG-16 models (with implicit weight normalization), one with standard training, which is not expected to be robust, and one with PGD-based adversarial training (Madry et al., 2018) with ℓ_∞ perturbations of budget $\epsilon = 2/255$, which is expected to be robust against a number of other perturbations as well (Yi et al., 2021). The HaH model in (Cekic et al., 2022) is used as another baseline for robustness which, like our approach, supplements training with layer-wise costs (the model they report on modifies 6 layers).

Model	Clean	Noise $\sigma = 0.1$	Avg/Min/Max corruptions	Avg/Min/Max severity level: 5	Autoattack ℓ_2 adv, $\epsilon = 0.25$	Autoattack ℓ_∞ adv, $\epsilon = 2/255$
VGG-16 Std	**92.5**	24.8	72.7/47.6/90.6	55.2/22.3/87.4	13.6	10.3
VGG-16 Adv	88.3	80.0	79.6/52.8/86.1	70.9/20.4/85.0	72.1	**72.2**
HaH	87.4	61.7	76.6/58.5/86.2	67.2/46.3/83.2	25.8	19.9
TEXP-1	88.3	68.4	79.6/69.7/88.1	71.8/48.3/87.7	39.4	27.6
TEXP-1 Adv	87.3	**82.7**	82.9/74.8/86.4	78.2/49.1/84.7	70.8	65.8
TEXP-1 BAL Adv	89.0	81.1	**84.1/78.6/88.2**	**79.2/56.9/86.2**	**75.1**	70.7

Table 1. Enhanced robustness to corruptions under VGG-16 based TEXP models on CIFAR-10 clean and corruptions datasets

Corruptions →	Noise				Weather					Blur					Digital				
Models ↓	Gauss.	Shot	Speck.	Imp.	Snow	Frost	Fog	Brig.	Spat.	Defoc.	Gauss.	Glass	Motion	Zoom	Cont.	Elas.	Pixel.	JPEG	Satur.
VGG-16 Std	24.6	32.9	39.9	22.1	73.9	61.8	**64.7**	**87.4**	68.1	49.6	39.3	48.0	60.7	61.0	22.3	75.6	56.3	77.7	82.2
VGG-16 Adv	80.1	81.1	79.7	62.6	75.1	74.1	32.5	77.9	78.0	72.2	67.8	77.1	69.6	75.9	20.4	79.0	83.0	85.0	76.8
HaH	61.7	61.7	59.2	46.3	73.8	72.3	62.8	83.2	76.7	64.3	58.4	53.2	65.1	68.9	76.0	74.0	60.5	79.3	79.6
TEXP-1	68.4	70.8	68.7	48.3	75.8	77.0	61.0	84.6	73.5	69.0	63.8	64.2	66.9	72.8	**87.7**	74.5	74.8	81.6	80.4
TEXP-1 Adv	**82.7**	**82.7**	**81.9**	**73.9**	76.6	81.5	49.1	81.2	80.4	76.9	**74.4**	77.7	75.3	79.2	84.0	79.3	**83.9**	84.7	81.1
TEXP-1 BAL Adv	81.2	81.5	80.8	69.9	**80.1**	**83.0**	56.9	84.2	**82.6**	**77.0**	74.2	**77.8**	**76.4**	**79.7**	86.2	**81.3**	83.7	**85.4**	**83.0**

Table 2. Robustness to common corruptions of the highest severity level in the CIFAR-10-C dataset

Our TEXP models. We modify only the first layer of the VGG-16 to a TEXP layer. The hyperparameters are set as follows: training tilt $t = 1$, inference tilt $t_{\text{inf}} = 0.1$, layer weight $\alpha = 0.0001$. We utilize the TEXP objective (13) for training. The parameters were chosen based on a coarse grid search for tilts and layer weight. Note that the tilt in the training objective is larger than that in the softmax layer. We also report on promising results combining TEXP with adversarial training, with ℓ_∞ perturbations of budget $\epsilon = 2/255$, using both the standard and balanced TEXP objectives (termed TEXP-1 Adv and TEXP-1 BAL Adv respectively). For simplicity, all hyperparameters for TEXP across our three models are set to those for the basic TEXP model (better performance may be obtained by further fine-tuning). As we shall see, the balanced TEXP objective combined with adversarial training provides the best results, but we note that it also requires more careful optimization: setting large layer weight α in the initial stages of training may degrade performance, since filters are initialized randomly and a strong demotion of weak mismatches may not be desirable at an early stage.

Training. The end-to-end discriminative cost is taken to be the cross-entropy loss. We employ the ADAM optimizer (Kingma & Ba, 2014) with a multi-step learning rate, beginning with 0.001, and decreasing by a factor of 10 at epochs 60 and 80. We train all models for 100 epochs.

Evaluation metrics. We evaluate over 19 different common corruptions on the CIFAR-10-C (Hendrycks & Dietterich, 2018) dataset. We report the average, minimum and maximum over all the corruptions, for both the entire dataset comprising of 5 different severity levels, and also on specifically the corruptions of the highest severity. We also separately report on the corrupted data formed by the addition of Gaussian noise with standard deviation $\sigma = 0.1$, since the motivation for our approach spans from estimation under Gaussian noise. In addition, we find that our approach provides robustness to mild adversarial perturbations ($\epsilon = 0.25$ for ℓ_2 and $\epsilon = 2/255$ for ℓ_∞ respectively). In this adversarial evaluation, we use AutoAttack (Croce & Hein, 2020), suggested by RobustBench (Croce et al., 2020), which is parameter-free and consists of a suite of different attacks (in particular, we employ the APGD-CE and APGD-T attacks).

Improvement in robustness against corruptions. Table 1 lists the test accuracies of the benchmark and TEXP models under different data distortions. TEXP-1 (a single TEXP layer, no data augmentation) provides gains in robustness to noise and other out-of-distribution (OOD) corruptions (both at the highest severity level and all levels) in comparison to standard VGG and HaH models. In comparison with adversarial training alone, TEXP-1 provides better OOD robustness against the worst corruption (i.e., the minimum accuracy among all corruptions). Compared to baseline VGG, TEXP-1 improves robustness to noise from 24.8% to 68.4%, OOD robustness averaged over all different types of corruptions from 55.2% to 71.8% for the highest severity level. It also provided increased robustness to mild adversarial perturbations.

TEXP (both the standard and balanced objectives) combined with adversarial training provides even more powerful enhancements in OOD robustness, outperforming both the HaH and the adversarial training benchmarks, both in terms of average over all kinds of corruptions and the minimum or worst-case among the different corruptions. Compared to adversarial training alone, the TEXP-1 BAL Adv im-

Model	Clean	Noise $\sigma = 0.1$	Avg/Min/Max corruptions	Avg/Min/Max severity level: 5	Autoattack ℓ_2 adv, $\epsilon = 0.25$	Autoattack ℓ_∞ adv, $\epsilon = 2/255$
ResNet-20 Std	90.0	21.5	66.8/38.8/87.6	49.5/21/83.9	0.6	0.2
ResNet-20 Adv	85.7	72.7	76.7/49.8/83.7	67.8/21.5/83	69.8	69.5
TEXP-6 ResNet-20	85.7	69.1	77.7/69.4/83.9	70.4/53.3/81.3	34.7	21.6
TEXP-6 ResNet-20 Adv	84.0	76.2	78.7/69.2/82.7	73.4/43.2/81.3	67.2	62.1

Table 3. Enhanced robustness to common corruptions under ResNet-20 based TEXP models on CIFAR-10 clean and corruptions datasets

proved OOD robustness from 79.6% to 84.1% (min accuracy from 52.8% to 78.6%) for all levels and from 70.9% to 79.2% (min accuracy from 20.4% to 56.9%) for the highest severity. It also provides comparable adversarial robustness against mild adversarial attacks.

Table 2 reports the robustness of the models to each of the 19 common corruptions separately for the highest severity level of 5, and shows that TEXP models are superior in obtaining robustness across the board. While vanilla adversarial training helps in robustness to noise, it deteriorates performance against corruptions like contrast, fog and brightness (Yin et al., 2019; Kireev et al., 2021; Machiraju et al., 2022). The TEXP based models remedy this remarkably for contrast, and alleviate this effect for fog and brightness.

Applicability to different architectures and deeper layers. We illustrate the broader applicability of TEXP via the ResNet-20 (He et al., 2016) model. This comprises a 3×3 convolution layer followed by 3 blocks, each containing 3 residual units. We modify the first 6 layers to TEXP layers by replacing the batch norm and leaky ReLU by tilted softmax and thresholding. The results, shown in Table 3, demonstrate improved OOD robustness despite minimal effort in hyperparameter tuning: the parameters used are $t = 5.0$, $t_{\text{inf}} = 0.1$, and $\alpha = 0.001$ for all the 6 layers, in both standard and adversarial TEXP models. We expect that more fine-grained adjustments of tilts for individual layers will further enhance performance, but these preliminary results do illustrate the potential gains from applying TEXP to multiple layers.

6. Conclusion

We have presented promising preliminary results indicating that the robustness of neural models can be enhanced by architectural modifications inspired by communication theory, supplementing end-to-end training with layer-wise TEXP objective functions, and replacing ReLU and batch norm by softmax and thresholding in the inference path. In order to compare with the benchmarks on layer-wise training set in (Cekic et al., 2022), we have focused on the VGG architecture with the CIFAR-10 dataset. We have demonstrated that even a single TEXP layer significantly improves OOD robustness against common corruptions without requiring data augmentation. Adversarial training with small perturbation budgets is also known to improve OOD robustness. We show that TEXP performance (without augmentation) against common corruptions is superior to that of adversarial training, while TEXP appropriately combined with adversarial training yields strong performance across the board against common corruptions.

As a quick check on the applicability of TEXP to different architectures and multiple layers, we also provide preliminary results on ResNet which show gains in robustness. We plan to build on these promising results in several directions, including more extensive experimentation for different datasets and architectures, development of communication-theoretically motivated guidelines for tuning of TEXP hyperparameters, and further exploration of combining TEXP with adversarial training and other simple augmentation techniques. Finally, while we have focused here on broad spectrum robustness, showing performance gains against common corruptions and mild adversarial attacks, an interesting direction for future work is to adapt our communication-theoretic approach for robustness against strong adversarial attacks.

Broader Impact & Limitations

Traditionally, robustness has been improved through the application of data augmentations and optimization of end-to-end costs. Our approach takes a different route by focusing on gaining more control over intermediate layer outputs and aligns with the broader goal of making deep networks more transparent. Furthermore, our method has the potential to work well with other data augmentation techniques, thereby expanding its applicability to various tasks.

A limitation of our current work is that we are yet to develop concrete guidance on setting the tilt parameters, which is useful to optimize the performance of our approach for different architectures. Furthermore, optimizing layer-wise costs for very large dimensional data is computationally intensive. We expect that future work will focus on maximizing the effectiveness of our approach across different datasets and models, to ascertain the generalizability of the TEXP approach. Nonetheless, our findings underscore the value of layer-wise tilted exponentials in enhancing robustness to OOD corruptions, which is important in many practical machine learning tasks where test samples in the real-world are often different from the curated training data.

References

Beirami, A., Calderbank, R., Christiansen, M. M., Duffy, K. R., and Médard, M. A characterization of guesswork on swiftly tilting curves. *IEEE Transactions on Information Theory*, 2018.

Butler, R. W. *Saddlepoint approximations with applications*. Cambridge University Press, 2007.

Calian, D. A., Stimberg, F., Wiles, O., Rebuffi, S.-A., Gyorgy, A., Mann, T., and Gowal, S. Defending against image corruptions through adversarial augmentations. *arXiv preprint arXiv:2104.01086*, 2021.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

Cekic, M., Bakiskan, C., and Madhow, U. Neuro-inspired deep neural networks with sparse, strong activations. In *ICIP*, pp. 3843–3847. IEEE, 2022.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 113–123, 2019.

Dembo, A. and Zeitouni, O. *Large deviations techniques and applications*. Springer Science & Business Media, 2009.

Dodge, S. and Karam, L. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, pp. 1–7. IEEE, 2017.

Gilmer, J., Ford, N., Carlini, N., and Cubuk, E. Adversarial examples are a natural consequence of test error in noise. In *International Conference on Machine Learning*, pp. 2280–2289. PMLR, 2019.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

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

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*, 2018.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=S1gmrxHFvB.

Huang, T., Halbe, S. A., Sankar, C., Amini, P., Kottur, S., Geramifard, A., Razaviyayn, M., and Beirami, A. Robustness through data augmentation loss consistency. *Transactions on Machine Learning Research*, 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kireev, K., Andriushchenko, M., and Flammarion, N. On the effectiveness of adversarial training against common corruptions. *arXiv preprint arXiv:2103.02325*, 2021.

Kireev, K., Andriushchenko, M., and Flammarion, N. On the effectiveness of adversarial training against common corruptions. In *Uncertainty in Artificial Intelligence*, pp. 1012–1021. PMLR, 2022.

Kort, B. W. and Bertsekas, D. P. A new penalty function method for constrained minimization. In *IEEE Conference on Decision and Control and 11th Symposium on Adaptive Processes*, 1972.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Li, T., Beirami, A., Sanjabi, M., and Smith, V. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021.

Li, T., Beirami, A., Sanjabi, M., and Smith, V. On tilted losses in machine learning: Theory and applications. *Journal of Machine Learning Research*, 24(142):1–79, 2023. URL http://jmlr.org/papers/v24/21-1095.html.

Liu, G.-H. and Theodorou, E. A. Deep learning theory review: An optimal control and dynamical systems perspective. *arXiv preprint arXiv:1908.10920*, 2019.

Machiraju, H., Choung, O.-H., Herzog, M. H., and Frossard, P. Empirical advocacy of bio-inspired models for robust

image recognition. *arXiv preprint arXiv:2205.09037*, 2022.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Pee, E. and Royset, J. O. On solving large-scale finite minimax problems using exponential smoothing. *Journal of Optimization Theory and Applications*, 2011.

Qin, Y., Wang, X., Lakshminarayanan, B., Chi, E. H., and Beutel, A. What are effective labels for augmented data? improving calibration and robustness with autolabel. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023.

Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., and Bethge, M. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33: 11539–11551, 2020.

Siegmund, D. Importance sampling in the monte carlo study of sequential tests. *The Annals of Statistics*, 1976.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Tack, J., Yu, S., Jeong, J., Kim, M., Hwang, S. J., and Shin, J. Consistency regularization for adversarial robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8414–8422, 2022.

Yi, M., Hou, L., Sun, J., Shang, L., Jiang, X., Liu, Q., and Ma, Z. Improved ood generalization via adversarial training and pretraing. In *International Conference on Machine Learning*, pp. 11987–11997. PMLR, 2021.

Yin, D., Gontijo Lopes, R., Shlens, J., Cubuk, E. D., and Gilmer, J. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.