MAXENT LOSS: CONSTRAINED MAXIMUM ENTROPY FOR CALIBRATING DEEP NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Miscalibration distorts our interpretation of a model's confidence and correctness, making it unreliable for real-world deployment. In general, we want meaningful probabilistic estimates of our model's *uncertainty*, which are essential in real-world applications. This may include inputs from out-of-distribution (OOD) data, which can be significantly different from the given training distribution. Motivated by the Principle of Maximum Entropy, we show that – compared to conventional cross-entropy loss and focal loss – training neural networks with additional statistical constraints can improve neural calibration whilst retaining recognition accuracy. We evaluate our method extensively on different augmented and in-the-wild OOD computer vision datasets and show that our MaxEnt loss achieves state-of-the-art calibration in all cases. Our code will be made available upon acceptance.

1 INTRODUCTION

Recent advances in machine learning have given rise to large neural networks with strong recognition performance in fields such as computer vision and natural language processing. They are increasingly used in areas where safety is a concern, such as self-driving cars (Bojarski et al., 2016), medical prognosis (Esteva et al., 2017; Bandi et al., 2019), and weather forecasting (Sønderby et al., 2020). Despite their popularity in these applications, deep neural networks have a tendency to be poorly calibrated. Calibration refers to the model's correctness with regards to its output probabilities that reflect its predictive uncertainty. In other words, models tend to misclassify samples with high confidence and erroneously recognize correct classes with low confidence. This leads to serious consequences, since the output probabilities of neural networks are overconfident, and their resulting decisions or actions cannot be trusted. Furthermore, a well-calibrated classifier should behave unconfidently and predict low probabilities whenever it misclassifies samples with high levels of uncertainty. This includes data samples which may be in-distribution (ID) but not present in the training set, or data from an out-of-distribution (OOD) set (Thulasidasan et al., 2019).

The current hypothesis as to why modern neural networks are poorly calibrated is that these large models with millions of parameters have the capacity to learn and overfit to the given training data (Guo et al., 2017). This is especially true if the model is trained using the cross-entropy (CE) loss (a.k.a. negative log-likelihood loss), because the CE loss is in the form of the upper bound of the Kullback-Leibler (KL) divergence which measures the statistical difference between the target and predicted distributions. For the multi-class classification problem, CE loss is only fully minimized when softmax probabilities $\hat{p}$ are equal to the onehot ground truth y such that ($\hat{p} = y$). This means that even though the accuracy is at 100%, CE loss can still be positive and minimized further by increasing the confidence of the predicted probabilities, resulting in overfitting and miscalibration. In general we want our models to not only remain accurate and well-calibrated in-distribution but to also provide further robustness against OOD shifts for the safe deployment of deep learning models (Amodei et al., 2016).

In this paper, we follow the works of (Mukhoti et al., 2020) to further explore the benefits of using the Principle of Maximum Entropy, also known as the MaxEnt method (Jaynes, 1957), and propose a novel loss function for improving model calibration based on constrained maximum entropy. Our method works by introducing additional constraints which complement loss functions typically used

in supervised learning. We provide systematic comparisons on model accuracy and calibration using image classification tasks. Our contributions can be summarized as follows:

1. **Constrained Maximum Entropy:** We show the link between the Principle of Maximum Entropy and Focal loss, exploring how introducing constraints can improve calibration.
2. **Novel loss formulation:** We propose MaxEnt loss with constraints for training well-calibrated models and compare it against CE loss, Focal loss and Poly loss.
3. **Evaluation on OOD shifts:** Our experiments show that MaxEnt loss remains robust in terms of recognition accuracy and model calibration for both augmented and in-the-wild distribution shifts. We further analyze its behavior under increasingly distorted inputs.
4. **Complements post-hoc calibration:** We further show how our method is nonrestrictive, and works well in unison with other post-hoc calibration methods such as temperature scaling.

2 Related Work

OOD shifts: For classification problems, we train a neural network to infer the posterior distribution and evaluate its performance on a test set that is ID with the given training set. For OOD problems, test samples do not align with the training samples, which can be caused by either: 1.) Completely OOD test inputs that belong to an OOD class not from any of the given ground truth labels (Du et al., 2022). 2.) Shifted OOD test inputs that may have been caused by illumination, perturbation shifts and corruptions (Hendrycks & Gimpel, 2017). For this work, we focus on shifted OOD from both augmentations and in-the-wild. For an in depth review of OOD shifts, we refer to (Wiles et al., 2022) for a recent summary of the literature.

Model calibration: Recent works have proposed various solutions for calibrating neural networks without losing recognition performance. A brief overview of the existing methods can be described as the following categories: 1.) Methods that learn to approximate the true joint distribution or latent hidden vector z using generative models such as those using Cycle-GAN (Zhu et al., 2017), VAE (Kingma & Welling, 2014). 2.) Methods that account for unknown classes not found during training time using an OOD-detection mechanism. (Du et al., 2022) 3.) Regularization techniques that directly manipulate the predicted output probabilities such as those using augmentations, architecture choice, isotonic regression (Zadrozny & Elkan, 2002), D-transforms (Parikh & Chen, 2008), dropout (Srivastava et al., 2014), label smoothing (Pereyra et al., 2017), Bayesian binning (Naeini et al., 2015) and temperature scaling (Guo et al., 2017).

Maximum Entropy: In contrast to other regularization methods, (Pereyra et al., 2017) have shown that directly penalizing neural networks with the maximum entropy confidence penalty term helps prevent overconfidence and peaked predictions, resulting in better generalization behavior. Another related work, Focal loss (Lin et al., 2017) was originally proposed for object detection and multi-label image classification problems, but helps improve overall calibration error in other classification problems. Mathematically, it can be shown that Focal Loss is a general form of CE loss with an addition confidence penalty term (Mukhoti et al., 2020). This is because reducing Focal loss simultaneously *minimizes* the KL divergence between the predicted probabilities and the target distribution (typically in the form of one-hot encoded vectors) and *maximizes* the Shannon entropy between these two distributions, therefore discouraging the model from being too overconfident with its predictions. Training neural models with Focal loss has been shown to include benefits such as reducing overfitting, improving generalization behaviour and reducing calibration error.

By itself, the maximum entropy confidence penalty term has deep connections to the Principle of Maximum Entropy (Jaynes, 1957), also commonly referred to as the MaxEnt method, which has a long standing in statistics and information theory. Using supervised learning, we can maximize the model's entropy subject to constraints on useful statistics observed in the training set (Berger et al., 1996). The MaxEnt method has also found its way into other computer vision tasks, such as Fine-Grained Visual Classification (FGVC) where classes may be visually similar to one another, making it harder to differentiate between samples (Dubey et al., 2018). In reinforcement learning, policies with high entropy output distributions tend to encourage stochasticity and improve exploration (Williams & Peng, 1991; Haarnoja et al., 2018).

3 METRICS AND METHODS

3.1 CALIBRATION METRICS

In theory, a model is considered *perfectly* calibrated if and only if the model's predicted probabilities match the true posterior distribution. Specifically ideal model calibration can be defined as if and only if: $P(\hat{Y} = y|\hat{p} = p) = p \quad \forall \in p[0-1]$, where Y represents the given ground truth, $\hat{Y}$ is the predicted class label and $\hat{p}$ are its associated predicted probabilities. Realistically, following this definition of calibration is impossible since the true posterior distribution remains unknown, therefore the following calibration error metrics have been proposed.

Expected Calibration Error (ECE) (Naeini et al., 2015) is a scalar quantity computed by splitting the model's predicted probabilities into equally spaced B number of bins, where B is a user-defined parameter; let N be the total number of samples and n_b be the total number of samples for each individual bin. acc represents the average accuracy and $conf$ is the average probability for each partitioned bin. The weighted absolute differences between the accuracy acc and confidence $conf$ for each partitioned bin is taken and averaged across all bins. Specifically, $ECE = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - conf(b)|$.

Maximum Calibration Error (MCE) (Naeini et al., 2015) holds a similar idea to the ECE, where approximations include binning the predicted probability scores and measuring the maximum absolute difference between the partitioned accuracy and confidence bins. The MCE is suited for high-risk applications where the worst case scenarios are to be considered carefully $MCE = \max_{\{1 \in \dots B\}} |acc(b) - conf(b)|$.

Brier Score (BS) (Brier, 1950; Degroot & Fienberg, 1983) is defined as the sum of squared errors between the predicted probabilities and the one-hot ground truth vector: $BS = \frac{1}{N} \sum_{n=1}^{N} |y_n - \hat{p}_n|^2$.

Negative Log-likelihood (NLL) (Hastie et al., 2001) , also commonly known as CE loss in deep learning (Goodfellow et al., 2016). Given a model's probabilities $\hat{p}_n$ and the groundtruth onehot vectors y_n, $NLL = -\frac{1}{N} \sum_{n=1}^{N} y_n \log(\hat{p}_n)$.

Generally, classifiers with lower overall ECE and MCE calibration scores are considered to be better calibrated. In the event that both ECE and MCE are equal to zero, the classifier is considered perfectly calibrated.

3.2 PRINCIPLE OF MAXIMUM ENTROPY, FOCAL LOSS AND POLY LOSS

We show the relationship between the Principle of Maximum Entropy (Jaynes, 1957), confidence penalty term (Pereyra et al., 2017) and Focal Loss (Lin et al., 2017), and how they are the unconstrained form of the MaxEnt method. The MaxEnt method is a probability distribution that satisfies a limited number of constraints by maximizing the Shannon entropy subject to all of the given constraints. Specifically, the general mathematical form of the MaxEnt method is given by:

$$\begin{aligned} &\text{maximize } H(p(\boldsymbol{y_n}|\boldsymbol{x_n})) = -\sum p(\boldsymbol{y_n}|\boldsymbol{x_n}) \log p(\boldsymbol{y_n}|\boldsymbol{x_n}) \\ &\text{subject to} \sum p(\boldsymbol{y_n}|\boldsymbol{x_n}) f_i(y) = c_i \text{ for all constraints} f_i(y) \end{aligned} \tag{1}$$

where $f_i(y)$ is a function of the random variable or class $\boldsymbol{y_n}$ for the nth input $\boldsymbol{x_n}$ and the different given constraints are represented by c_i and the predicted conditional distribution $p(\boldsymbol{y_n}|\boldsymbol{x_n})$. As a mandatory constraint we require the output probability scores from the classifier to be valid probabilities such that $\sum p(\boldsymbol{y_n}|\boldsymbol{x_n}) = 1$. This can be easily fulfilled by using either the softmax or sigmoid activation function. With the valid probability constraint, the general solution to the MaxEnt problem can be written as:

$$p(\boldsymbol{y_n}|\boldsymbol{x_n}) = \frac{1}{Z} e^{\sum -\lambda_i f_i(y)} \tag{2}$$

where $Z = e^{1+\lambda_0}$ is also commonly known as the partition function, and λ_i are the respective Lagrange multipliers for each constraint.

Next, we show the relation to Focal loss, which was originally designed to allow the model to concentrate on harder samples and reduce the emphasis on easily classified samples (Lin et al.,

2017). Consider, the multi-class form of the Focal loss where γ is a user defined hyperparameter. Notice, that by setting $\gamma = 1$, the Focal loss can be expanded and re-written as the CE loss with an additional confidence penalty term (Shannon's entropy term):

$$\begin{aligned}\mathcal{L}_F &= -\sum (1 - p(\boldsymbol{y_n}|\boldsymbol{x_n}))^{\gamma} \log p(\boldsymbol{y_n}|\boldsymbol{x_n}) \\ &= -\sum \underbrace{\log p(\boldsymbol{y_n}|\boldsymbol{x_n})}_{\text{CE loss}} - \underbrace{H(p(\boldsymbol{y_n}|\boldsymbol{x_n}))}_{\text{Shannon term}}\end{aligned} \tag{3}$$

In the case where other values of γ are chosen such that $\gamma > 1$, the general form of Equation 3 still holds true such that the Shannon entropy term is accompanied by another polynomial. Since, further minimizing the CE loss on correctly classified samples with high confidences will lead to overfitting and result in miscalibration (Guo et al., 2017), the additional entropy term is useful for preventing peaked distributions and leads to better overall generalization performance (Mukhoti et al., 2020).

Another alternative to Focal loss is Poly loss (Leng et al., 2022), which differs from Focal loss by adjusting the loss curve vertically instead of horizontally. Poly loss uses hyper-parameters ϵ and j to perturb the leading polynomial coefficients and is defined as follows:

$$\mathcal{L}_P = -\sum \underbrace{\log p(\boldsymbol{y_n}|\boldsymbol{x_n})}_{\text{CE loss}} + \underbrace{\epsilon(1 - p(\boldsymbol{y_n}|\boldsymbol{x_n}))^j}_{\text{Poly term}} \tag{4}$$

3.3 Constrained MaxEnt Loss

In general, the MaxEnt method works without using additional knowledge about the problem apart from maximizing the entropy and the given constraints. Using the method alone would result in the best estimations of the target probability distribution with the highest levels of uncertainty which would lead to the model's estimation to be minimally biased towards unseen OOD data (Jaynes, 1957). This would mean that the MaxEnt method relies heavily on the given statistical constraints and does not consider other natural laws which may be inherent in the dataset such as the pixel distribution of an image dataset.

Therefore, we propose to train ours models using a general loss function such as Focal loss, which minimizes CE loss and maximizes the entropy as per the MaxEnt method. We can further constrain Focal loss by considering statistical constraints observed in the distribution of the training set. We mainly focus on the following two constraints, which can be easily added to the Focal loss.

Mean Constraint: If the value of the expected average μ of the target distribution is known, we can add the following terms to Equation 3:

$$\mathcal{L}^E_{ME} = \mathcal{L}_F + \lambda_0 \underbrace{(\sum p(\boldsymbol{y_n}|\boldsymbol{x_n}) - 1)}_{\text{Valid probability constraint}} + \lambda_1 \underbrace{(\sum \boldsymbol{y_n} p(\boldsymbol{y_n}|\boldsymbol{x_n}) - \mu)}_{\text{Mean constraint}} \tag{5}$$

where the first term represents the valid probability constraint and the second term represents the mean constraint. We use λ_1 as the Lagrange multiplier for the mean constraint which is set to $\lambda_1 = -\log \frac{\mu}{1+\mu}$ and can be computed from the prior distribution of the training set.

Variance Constraint: Consequently, if both the expected average μ and variance σ^2 of the target distribution are known, then the variance constraint is added to Equation 3 instead of the mean constraint. The Appendix contains proofs for each Lagrange multiplier and that for the case of the variance constraint $\lambda_1 = 0$. Which implies that the mean constraint can be removed since it is already intrinsically considered in the definition of the variance constraint with $\lambda_2 = \sigma^{-2}$. Therefore, the final objective function to be minimized is given by the following Lagrangian:

$$\mathcal{L}^G_{ME} = \mathcal{L}_F + \lambda_0 \underbrace{(\sum p(\boldsymbol{y_n}|\boldsymbol{x_n}) - 1)}_{\text{Valid probability constraint}} + \lambda_2 \underbrace{(\sum (\boldsymbol{y_n} - \mu)^2 p(\boldsymbol{y_n}|\boldsymbol{x_n}) - \sigma^2)}_{\text{Variance constraint}} \tag{6}$$

To minimize either mean or variance constraint, we perform regression and train our models with the rest of the loss function as per Equation 5 or Equation 6. For our work, we make use of the L2 loss for the constraints and their respective optimal Lagrange multipliers. We further show how the above constraints result in the Exponential and Gaussian distributions respectively in the Appendix.

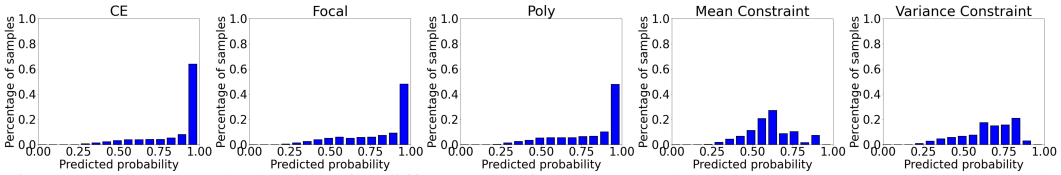

Figure 1: Bin-strength densities for different loss functions, evaluated on the augmented OOD FashionMNIST validation set.

We compare the bin-strength plots of different loss functions evaluated on the OOD FashionMNIST validation set, highlighting the effects of including additional constraints using our proposed method. With CE loss, the predicted probabilities are most confident and have the "peakiest" distribution. With Focal loss or Poly loss, the predictions are slightly reduced and better calibration can be achieved.

In contrast, by adding the mean and variance constraints during training, our method produces significantly "softer" output probabilities as compared to "peakier" predictions made by the model when trained with CE loss, Focal loss and Poly loss. Our method is able to further smoothen the probability distribution and achieve better calibration error without losing any recognition accuracy as shown later in our experiment results.

4 Experiments and Results

We evaluate the performance of our proposed method against other baselines using image classification experiments. Firstly, we briefly describe all the augmented and wild OOD datasets used in our experiments, secondly we compare the MaxEnt loss against CE loss, Focal loss and Poly loss and additionally evaluate our results on increasingly shifted data. Finally, we compare overall performance gains and post-hoc calibration behaviours.

4.1 Datasets

Augmented OOD: We make use of image transformations to replicate the possible shifts in OOD data. We use the following four computer vision datasets and augment their validation and test sets.

1. MNIST (LeCun et al., 2010): Handwritten digits of (28x28) grayscale images consisting of ten different classes, with 45,000/15,000/10,000 samples for training/validation/testing.
2. FashionMNIST (Xiao et al., 2017): A drop-in alternative to MNIST, consisting of ten classes of (28x28) grayscale fashion images, with 45,000/15,000/10,000 examples for training/validation/testing.
3. CIFAR10/CIFAR100 (Krizhevsky & Hinton, 2009): RGB colored images (32x32) with ten or hundred classes. 40,000/10,000/10,000 images for training/validation/testing.

Wild OOD: We use the following three wild computer vision datasets with their provided ID training sets and OOD sets for validation and testing.

1. Camelyon17-Wilds (Bandi et al., 2019): Binary classification task on whether a (32x32) tissue slide contains any malignant/benign tumours.
2. iWildCam-Wilds (Beery et al., 2020): Static camera traps deployed across different terrains with radical shifts in camera pose, background and lighting. Task is to identify the species in the photo out of 182 animal classes.
3. FMoW-Wilds (Christie et al., 2018): Satellite images across different functional buildings and lands from over 200 countries. The task is to detect one out of 62 categories, including a "false detection" category.

The datasets used for our experiments are shown in Figure 2 of the Appendix. The validation/test sets used in augmented OOD are shifted in levels of increasing difficulty. For our data augmentation strategy, we follow a routine similar to (Deng & Zheng, 2021), where the training set is kept minimally modified with augmentations only applied to the validation and test sets. For our augmentations, we randomly create four tuples each consisting of different transformations, with T0 minimally modified and T3 containing the most transforms. We select a variety of standard image transformations such as flipping, rotations, affinities, perspectives, colour jitter, sharpening and

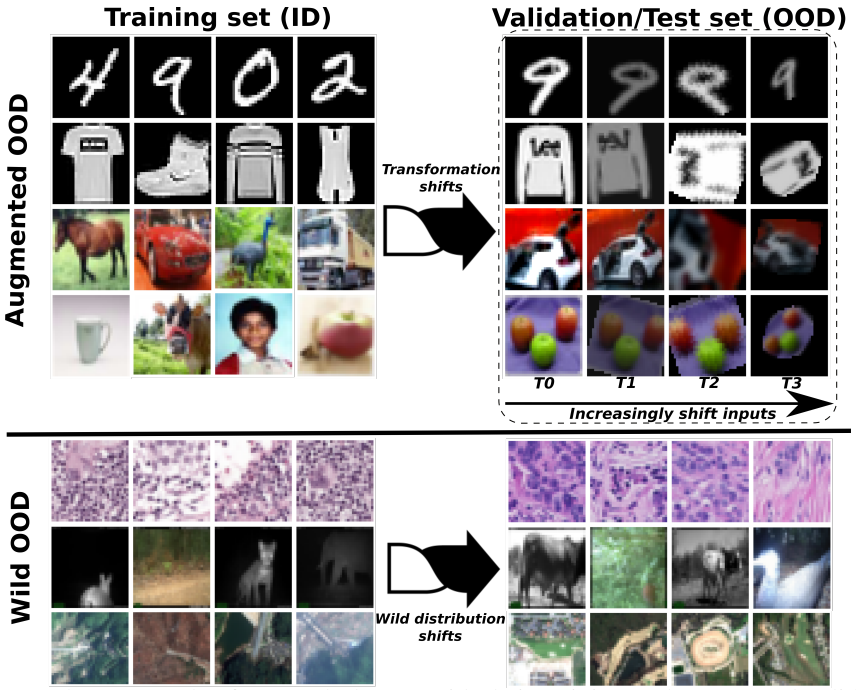

Figure 2: We show examples from each dataset with their training and augmented validation/test images for MNIST, FashionMNIST, CIFAR10 and CIFAR100 respectively. For wild OOD, we show examples from Camelyon17-Wilds, iWildCam-Wilds and FMoW-Wilds datasets.

Gaussian blurs in magnitudes of increasing difficulty. Specifically, T0 is minimally shifted such that the distributions of the training, validation and test sets are aligned and that T3 contains the most transformations for the augmented validation and test sets.

We find this augmentation strategy to be useful in our experiments, since the total number of augmented test sets and varying difficulties can be easily controlled. We denote real-world OOD datasets with "Wilds" as per (Koh et al., 2021) and use pretrained ResNet-18 (He et al., 2016) and DenseNet-121 Huang et al. (2017) for image classification tasks. For our analysis, we use Adam optimizer (Kingma & Ba, 2015), with a constant learning rate of 2.5e-4, and train 200 epochs with a batch size of 1024 for augmented OOD datasets and 50 epochs with a batch size of 256 for wild OOD datasets. Additional details regarding the transformations and experiments used can be found in the Appendix.

4.2 Illustrated Example - FashionMNIST and CIFAR-10

We illustrate our experiments using the FashionMNIST and CIFAR10 dataset and describe our findings below in Figure 3. For our comparisons, we use the probabilities from the softmax function after the fully connected layer as the model's measure of uncertainty and minimize different objective functions namely CE loss, Focal loss $\gamma = 1$, Poly loss $\epsilon = 1$ and our method with constraints placed on the mean and variance. We plot the training and validation accuracies as well as the validation ECE, MCE, Brier scores of the T1 augmented OOD set, we also show the learnt feature norms and discuss our findings below.

Expected Behavior: As anticipated, we expect to obtain high recognition on the training set, with all methods converging to a high accuracy score near 100% since training samples are kept simple and not modified with transformations. On the other hand, since the validation set is from a distribution shifted from the training set, we expect to see weaker recognition accuracy during inference. We highlight that a well-calibrated model should remain robust even under the influence of shifted OOD data, with higher levels of uncertainty and entropy. That is to say, a model should not be too confident in its predictions when it comes to samples that it is unfamiliar with and should remain maximally uncertain, especially when inputs are deviated from the learned training distribution.

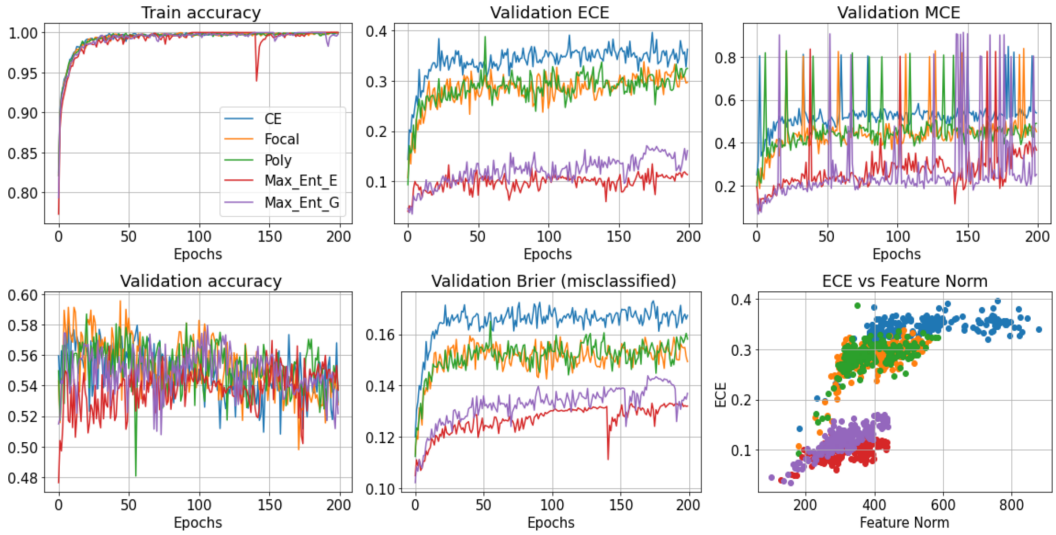

Figure 3: Metrics highlighting the performance on OOD FashionMNIST using ResNet-18 backbone architecture trained with different loss functions.

Empirical Observations: We see in Figure 3, that the models trained using the MaxEnt loss provide broadly better calibration scores, despite having similar accuracy scores in the training and validation sets. Clearly, solely using the validation accuracy to judge the performance of our models is insufficient and difficult to distinguish the model's ability to generalize and perform well on OOD samples.

Overall, we can clearly see significant improvements in the ECE, MCE and Brier score, with CE loss having the poorest calibration performance. We also plot the ECE versus the L2 norm of learnt the features, using the logits from the fully connected layer before the softmax layer. In general, most methods tend to have increasingly higher feature norms even after achieving the maximum accuracy. We further observe the correlation between the ECE and the feature norms, displaying a general trend that models with higher feature norms have higher expected calibration errors. Comparing our method against other baselines, models trained with MaxEnt Loss have lower calibration errors and smaller feature norms.

Overall Performance Gains: We report the performances on the OOD test sets, namely the accuracy, ECE, MCE (computed using 10 bins), NLL and Brier scores of misclassified samples in Table 1 along with their post temperature scaling performance. For the experiments shown in this table, we mainly used the set of T1 transforms for augmented OOD and the corresponding test sets for wild OOD. Firstly, we find that most loss functions produce relatively similar recognition accuracies across all datasets regardless of augmented or wild OOD. Secondly, it is clear that models trained with additional constraints using the MaxEnt loss outperforms all other baselines in terms of the calibration metrics along with competitive classification accuracies.

Another important finding is that MaxEnt loss is able to consistently deliver well calibrated models, even without the use of any additional ad-hoc calibration techniques such as temperature scaling. We are not able to find any major differences yet between the mean constraint versus the variance constraint form of the MaxEnt loss, with perhaps only slightly better performance coming from the variance constraint. Where the final performance of each loss function would still ultimately depend on the distribution of the given dataset. In addition to improved calibration performance, we further discuss the post-hoc calibration performances below.

Post-hoc Calibration: We choose temperature scaling as our post-hoc calibration technique, which linearly scales the classifier's output logits with a scalar T where $T > 0$. We follow the recommendations of (Mukhoti et al., 2020) and perform grid-search over a standard range of temperature values such as [1.25, 1.50, 1.75, 2.00], picking the optimal temperature which maximizes the validation set accuracy. If we consider the case for ad-hoc calibration, our method still manages to broadly produce the lowest calibration errors even after temperature scaling with statistically significant differences in all calibration metrics apart from the Brier score.

	Method	CE loss		Focal loss		Poly loss		**MaxEnt Mean**		**MaxEnt Var**	
		Pre-T	Post-T	Pre-T	Post-T	Pre-T	Post-T	Pre-T	Post-T	Pre-T	Post-T
MNIST	Accuracy	**0.674**	**0.678**	0.660	0.667	0.667	0.659	0.659	0.660	0.655	0.662
	ECE	0.223	0.234	0.254	0.154	0.233	**0.073**	0.111	0.138	**0.087**	0.182
	MCE	0.473	0.468	0.517	0.324	0.515	**0.197**	0.316	0.228	**0.190**	0.286
	Brier	0.157	0.158	0.165	0.140	0.160	0.124	**0.132**	0.129	0.136	**0.109**
	NLL	5.127	4.914	5.561	3.679	5.180	3.061	4.808	4.483	**4.585**	**3.037**
Fashion-MNIST	Accuracy	0.555	0.545	0.557	0.548	**0.565**	0.536	0.551	0.537	0.543	**0.549**
	ECE	0.323	0.245	0.272	0.098	0.248	0.136	**0.085**	0.081	0.132	**0.101**
	MCE	0.481	0.345	0.411	0.148	0.372	0.205	**0.196**	**0.158**	0.209	0.173
	Brier	0.161	0.142	0.148	0.114	0.145	0.119	**0.128**	0.118	0.133	**0.123**
	NLL	5.590	3.882	4.168	2.650	**4.033**	**2.789**	4.033	3.449	4.394	3.642
CIFAR10	Accuracy	0.551	0.559	0.553	0.540	0.557	0.547	0.535	0.546	**0.575**	**0.573**
	ECE	0.360	0.318	0.327	0.255	0.323	0.252	**0.098**	**0.048**	0.115	0.050
	MCE	0.512	0.453	0.448	0.350	0.443	0.805	**0.260**	0.295	0.220	**0.139**
	Brier	0.171	0.160	0.162	0.142	0.161	0.142	**0.123**	**0.111**	0.137	0.117
	NLL	7.048	5.137	5.396	3.793	5.268	3.678	**3.902**	**3.098**	5.363	3.471
CIFAR100	Accuracy	0.259	**0.272**	0.264	0.268	**0.266**	0.267	0.242	0.246	0.244	0.247
	ECE	0.257	0.445	0.485	0.287	0.460	0.317	0.199	**0.043**	**0.172**	0.148
	MCE	0.419	0.599	0.649	0.447	0.605	0.500	0.273	**0.261**	**0.224**	0.464
	Brier	0.012	0.015	0.015	0.013	0.015	0.013	**0.012**	**0.010**	**0.012**	**0.010**
	NLL	**4.800**	7.026	8.102	4.945	7.575	5.210	5.261	**3.866**	5.320	3.789
Camelyon17-Wilds	Accuracy	**0.824**	**0.893**	0.559	0.712	0.733	0.564	0.684	0.785	0.665	0.789
	ECE	0.128	0.045	0.193	0.177	0.218	0.354	**0.063**	0.113	0.127	**0.044**
	MCE	0.244	0.126	0.321	0.289	0.286	0.372	**0.171**	**0.147**	0.190	0.179
	Brier	0.804	0.650	0.523	0.664	0.8564	0.758	0.628	**0.412**	**0.505**	0.494
	NLL	3.686	2.082	1.384	2.047	4.310	2.711	1.619	**1.035**	**1.253**	1.239
iWildCam-Wilds	Accuracy	0.378	0.303	0.350	**0.397**	**0.408**	0.359	0.387	0.320	0.343	0.377
	ECE	0.337	0.408	0.332	0.206	0.389	0.332	**0.125**	0.184	0.128	**0.142**
	MCE	**0.489**	0.575	0.617	0.309	0.561	0.506	0.802	0.881	0.718	**0.470**
	Brier	0.008	0.008	0.008	0.006	0.009	0.007	**0.006**	0.006	0.007	**0.004**
	NLL	5.735	4.625	5.615	3.670	7.506	4.815	**3.782**	3.171	4.550	**2.616**
FMoW-Wilds	Accuracy	**0.451**	**0.449**	0.415	0.441	0.190	0.183	0.361	0.363	0.376	0.339
	ECE	0.230	0.226	0.290	0.188	0.531	0.425	0.108	**0.026**	**0.021**	0.061
	MCE	0.368	0.341	0.415	0.287	0.737	0.616	0.165	**0.111**	**0.142**	0.170
	Brier	0.022	0.021	0.021	0.020	0.023	0.019	0.019	**0.016**	**0.017**	0.017
	NLL	4.739	4.448	4.187	3.895	4.039	**3.264**	5.015	3.356	**3.482**	4.239

Table 1: Test scores on the OOD datasets computed across different approaches for both pre- and post-temperature scaling, with the best scores highlighted in bold.

Our findings are particularly encouraging, since our method complements temperature scaling and does not restrict users to perform ad-hoc calibration should they choose to do so. We note that both constraints have similar performances in terms of calibration behaviour and believe that choosing when to use which constraint may likely depend on the nature of the training set. For further results and behaviours when deployed on different architectures, please refer to the Appendix.

Robustness against increasing difficulty: Unlike Focal loss which requires the tuning of a hyper-parameter γ on a separate validation set, our method does not require additional tuning, since the Lagrange multipliers λ_i can be computed before training. We further compare our method against CE and Focal loss tuned with γ set to 0, 1, 2 or 3, using Densenet-121 evaluated on CIFAR10. We show in Figure 4, the bar plots of the ECE, MCE and NLL by taking the mean and standard deviation across 100 epochs for the different methods on four different levels of difficulty T0 to T3.

We notice that most methods behave relatively well-calibrated under ID settings (since T0 is minimally shifted), which would help suggest that most loss functions would perform well if the distribution of the training set is aligned with the validation/test set. However, as the distribution progressively shifts, the predictive uncertainty of all methods start degrading with the poorest overall performance coming from CE loss. Which means that, models trained with only CE loss may easily become overconfident and miscalibrated when inputs vary greatly from the training set.

In contrast, our method remains relatively robust even under increasingly shifted inputs beating the other baseline methods by a clear margin across all calibration metrics, with slightly better performance coming from the variance constraint form of the MaxEnt loss (orange).

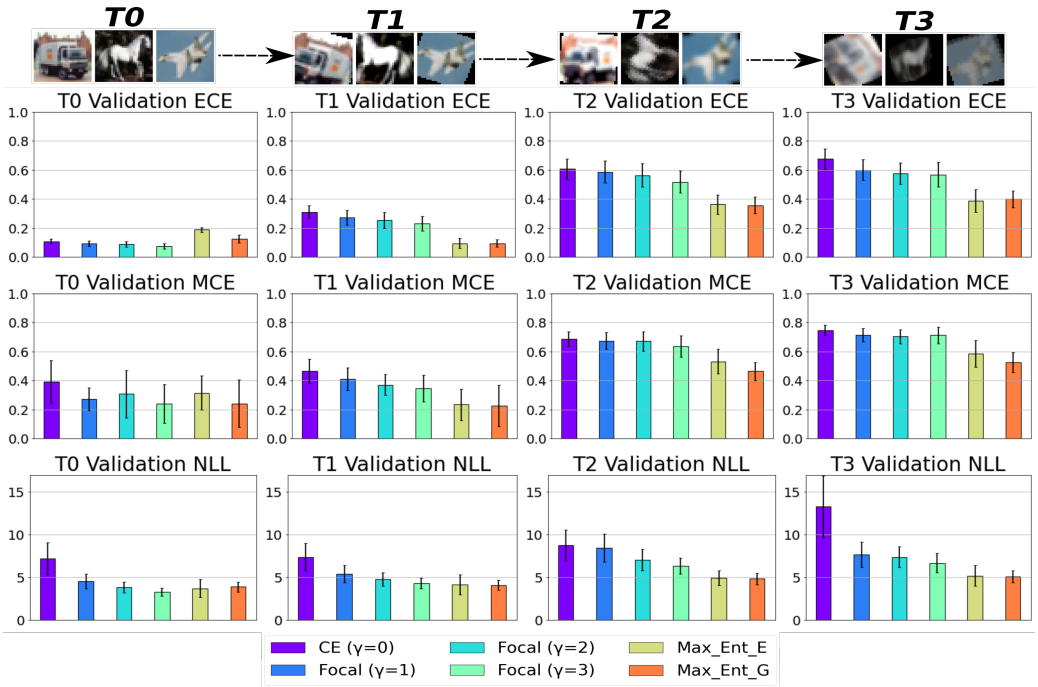

Figure 4: Bar plots for OOD CIFAR-10: Each row shows the calibration performances ECE, MCE & NLL degrading under increasingly shifted OOD inputs for each column left to right, T0 to T3 respectively.

5 Conclusions

We presented a novel training method in the form of a loss function for calibrating deep neural models in the face of OOD scenarios, across both forms of augmented and wild OOD for computer vision classification tasks. Our takeaway messages are as follows:

- Accuracy, by itself is not enough to measure the prediction quality of our models. Evaluating other metrics such as those presented in this paper such as the ECE/MCE/NLL/Brier scores are helpful in determining which models should be used for deployment.
- We show the relationship between Focal loss and the Principle of Maximum Entropy and propose a novel loss function using constraints for improving model calibration.
- MaxEnt loss does not require the use of validation set for hyperparameter tuning and outperforms other loss functions on both synthetic increasingly shifted OOD via image transformations and OOD in the wild.
- Well-calibrated models should perform well in both ID and OOD datasets. Post-hoc calibration methods that require validation set are calibrated on ID test sets but may not be calibrated on OOD.
- We show that MaxEnt loss is able to complement most post-hoc calibration methods such as temperature scaling and outperforms other loss functions pre- and post-tuning.
- Predictive uncertainty worsens under increasing dataset shift for most methods, whereas MaxEnt loss remains relatively robust without any additional post-hoc calibration.

Our loss function outperforms other loss functions such as CE loss, Focal loss and Poly loss on computer vision classification tasks with no significant increase in computation costs and requires only a few additional lines of code. We show how our method performs on both augmented and wild OOD datasets considering both pre- and post-hoc calibration. We also show comparisons on the behaviours of different loss functions subjected to shifted inputs, with our method outperforming other methods. In general, using the appropriate loss function can lead to well calibrated models, which in turn can help in the practical and trustworthy deployment of models since users can be notified when predictions may have too much uncertainty, improving reliability and fairness in AI. We hope that the MaxEnt loss is useful to the community and that our work provides better calibrated

models that can perform robustly even under distribution shifts, which remains a challenging hurdle for AI deployment.

REFERENCES

Dario Amodei, Christopher Olah, Jacob Steinhardt, Paul Francis Christiano, John Schulman, and Dandelion Mané. Concrete problems in AI safety. *arXiv*, abs/1606.06565, 2016.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019.

Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv*, abs/2004.10340, 2020.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

Mariusz Bojarski, David W. del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *arXiv*, abs/1604.07316, 2016.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6172–6180, 2018. doi: 10.1109/CVPR.2018.00646.

Morris H. Degroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32:12–22, 1983.

Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: Learning what you don't know by virtual outlier synthesis. *arXiv*, abs/2202.01197, 2022.

Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine-grained classification. In *NeurIPS*, 2018.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin M. Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep Learning*, volume 1. MIT Press, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proc. 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330, 2017.

Tuomas Haarnoja, Aurick Zhou, P. Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*, 2018.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, USA, 2001.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv*, abs/1610.02136, 2017.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, May 1957. doi: 10.1103/PhysRev.106.620.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv*, abs/1312.6114, 2014.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *Proc. 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664, July 2021.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto, 2009.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Zhaoqi Leng, Mingxing Tan, Chenxi Liu, Ekin Dogus Cubuk, Jay Shi, Shuyang Cheng, and Dragomir Anguelov. Polyloss: A polynomial expansion perspective of classification loss functions. In *Proc. International Conference on Learning Representations (ICLR)*, 2022.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *arXiv*, abs/1708.02002, 2017.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip HS Torr, and Puneet K Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 2020.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proc. 29th AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015.

Devi Parikh and Tsuhan Chen. Bringing diverse classifiers to common grounds: dtransform. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3349–3352, 2008. doi: 10.1109/ICASSP.2008.4518368.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv 1701.06548*, 2017.

Casper Kaae Sønderby, Lasse Espeholt, Jonathan Heek, Mostafa Dehghani, Avital Oliver, Tim Salimans, Shreya Agrawal, Jason Hickey, and Nal Kalchbrenner. Metnet: A neural weather model for precipitation forecasting. *arXiv*, abs/2003.12140, 2020.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya, and Sarah Ellen Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *NeurIPS*, 2019.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. *arXiv*, abs/2110.11328, 2022.

Ronald Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3:241–, 09 1991. doi: 10.1080/09540099108946587.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv*, abs/1708.07747, 2017.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2002. doi: 10.1145/775047.775151.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.