Triple-Hybrid Energy-based Model Makes Better Calibrated Natural Language Understanding Models

Haotian Xu
Ailbaba Group
htxu91@gmail.com

Yingying Zhang
East China Normal University
yyzhang@fem.ecnu.edu.cn

Abstract

Though pre-trained language models achieve notable success in many applications, it's usually controversial for over-confident predictions. Specifically, the in-distribution (ID) miscalibration and out-of-distribution (OOD) detection are main concerns. Recently, some works based on energy-based models (EBM) have shown great improvements on both ID calibration and OOD detection for images. However, it's rarely explored in natural language understanding tasks due to the non-differentiability of text data which makes it more difficult for EBM training. In this paper, we first propose a triple-hybrid EBM which combines the benefits of classifier, conditional generative model and marginal generative model altogether. Furthermore, we leverage contrastive learning to approximately train the proposed model, which circumvents the non-differentiability issue of text data. Extensive experiments have been done on GLUE and six other multiclass datasets in various domains. Our model outperforms previous methods in terms of ID calibration and OOD detection by a large margin while maintaining competitive accuracy.

1 Introduction

Since many industrial applications involve safety -critical domains such as healthcare (Li et al., 2019; Blinov et al., 2020; Li et al., 2020; Rasmy et al., 2021; Sarabadani, 2019), anticipating credit card defaults (Sun and Vasarhalyi, 2021) and self-driving (Khaitan et al., 2021), it's essential for machine learning systems to provide not only accurate but also well-calibrated predictions (Li et al., 2019), which can help to decide whether it can be trusted.

However, models achieving high accuracy usually lead to overconfidence and miscalibration (Guo et al., 2017; Thulasidasan et al., 2019; Ovadia et al., 2019). This motivates an interesting and important area that attempts to achieve a better trade-off between accuracy and calibration. In addition to ID calibration, it's more important for machine learning models to produce high uncertainty when OOD data is observed, rather than to produce wrong yet wildly confident predictions.

Related works. To overcome the problem of miscalibration, numerous methods have been proposed. The natural way is post-hoc calibration that transforms the output of the original network into calibrated confidence scores while maintaining the network's accuracy (Guo et al., 2017; Rahimi et al., 2020; Jung et al., 2020). The second method to mitigate miscalibration is to add regularizations during training such as label smoothing (Wang et al., 2020), Mixup (Zhang et al., 2018). Desai and Durrett (2020) and Kong et al. (2020) further conveys that the aforementioned methods can be applied to improve the calibration of pre-trained language models on NLU tasks. The third way is to design a specific loss function to minimize the discrepancy between accuracy and confidence. For example, Kong et al. (2020) lately propose the ID and OOD regularizer to leverage the relationship between accuracy and uncertainty, and it obtains a significant improvement over previous methods in ID calibration and OOD detection.

Energy-based Models. In another line of work, Joint EBM (JEM; Grathwohl et al., 2019) has been shown great improvements on ID calibration and OOD detection for images without explicit calibration correction mechanism. The core idea is to reinterpret a joint distribution $p_\theta(x, y)$ from a neural classifier $p_\theta(y|x)$ in the perspective of EBMs and jointly optimize the marginal distribution $p_\theta(x)$ and a neural classifier $p_\theta(y|x)$. Elflein et al. (2021) further investigate the OOD detection performance with different training approaches for $p_\theta(x)$ such as Stochastic Gradient Langevin Dynamics (SGLD; Welling and Teh, 2011), Sliced-Score-Matching (SSM; Song et al., 2020) and Variational Entropy Regularized Approximate maximum likelihood (VERA; Duvenaud et al., 2021).

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 274–285
May 2-6, 2023

Besides, Du and Mordatch (2019) propose an implicit generative models based on EBMs (IGEBM) and apply SGLD to optimize $p_\theta(x|y)$. It performs significantly better OOD detection than other generative models. However, as shown by Grathwohl et al. (2019), the accuracy of IGEBM has dropped dramatically to 49.1% on CIFAR10 while standard finetuning can achieve 95.8% accuracy. This result indicates that different loglikelihood factorization leads to great gaps in accuracy, ID calibration and OOD detection. Moreover, these training methods such as SGLD, SSM, VERA need to calculate the gradients about inputs, the none differentiability of text data limits the application of these methods on both calibration and OOD detection for NLU tasks.

Recently, He et al. (2021) proposes a joint training of classifier $p_\theta(y|x)$ and marginal distribution $p_\theta(x)$ based on Residual EBM (Deng et al., 2019) for NLU tasks. Different from JEM, their model is more flexible by designing various energy functions for marginal distribution without any restriction on joint distribution $p_\theta(x, y)$. To estimate the parameters of marginal distribution $p_\theta(x)$, they propose to apply noise contrastive estimation (NCE; Gutmann and Hyvärinen, 2010) to train the energy model by discriminating the real data and the fake data generated by a noise distribution. To make the noise distribution as close as possible to the data distribution, they finetune a task-specific GPT-2 (Radford et al.). Though it achieves improvements on ID calibration, it's often resource-intensive compared to previous methods to finetune GPT-2 (Li et al., 2022). Moreover, the quality and quantity of fake samples generated by noise distribution has great impacts for NCE training (He et al., 2021; Gutmann and Hyvärinen, 2010).

Contribution. Methodologically, we propose a novel model namely Triple-Hybrid Energy-based Model (THEM) based on the JEM (Grathwohl et al., 2019) through different decompositions of $\log p(x, y)$ into a unified framework. Compared to Grathwohl et al. (2019) and Du and Mordatch (2019), our model combines the classifiers $p(y|x)$, class-conditional density $p(x|y)$ and unconditional data density $p(x)$ into a hybrid model. Due to the none differentiability of text data, we further propose to adopt InfoNCE (Oord et al., 2018) with memory bank (He et al., 2020) to approximate the normalized constant of EBM efficiently. This method makes it possible for EBM training on NLU tasks which is not well explored in previous works regardless of input differentiability. We conduct comprehensive experiments with BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019) as the backbone and demonstrate the effectiveness of our framework on various datasets including GLUE (Wang et al., 2018) and six multi-class classification datasets (Kong et al., 2020) on various domains. Not only the experimental results show that our method achieves significant improvements in ID calibration and OOD detection with competitive accuracy over previous methods, but also it is more robust with respect to the temperature and size of memory bank compared to contrastive learning trained EBM including JEM(CL), IGEBM(CL) and HDGE.

Overall, the contributions can be summarized as follows:

- We propose a Triple-Hybrid Energy-based model (THEM) and apply InfoNCE with memory bank to optimize it efficiently and effectively for **discrete data**. It achieves significantly better performance compared to strong baselines including He et al. (2021) and Kong et al. (2020) in terms of ID calibration and OOD detection.
- We apply this training technique to JEM and IGEBM to obtain JEM(CL) and IGEBM(CL) respectively. THEM and JEM(CL) achieves better ID calibration and OOD detection compared to HDGE and IGEBM(CL) in average.
- We further study the effect of the temperature and size of memory bank for contrastive learning on ID calibration and OOD detection. THEM is more robust to these hyperparameters than JEM(CL) and HDGE(CL).

2 Preliminaries: Joint Energy Model and Contrastive Learning

Joint Energy Model (JEM). Energy-based models (EBMs; LeCun et al., 2006) measure the compatibility of the input variables $x \in \mathcal{X}$ and target variables $y \in \mathcal{Y}$ with an energy function $E_\theta(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which is the main building block. Low energy corresponds to high compatability. With E_θ, the probability for data in an EBM can be written as

$$p_\theta(x, y) = \frac{\exp\left(-E_\theta(x, y)\right)}{Z_\theta}, \tag{1}$$

where Z_θ is the normalizing constant. EBMs are flexible to parameterize since they do not make restrictions on the tractability of Z_θ.

Joint Energy Model (JEM; Grathwohl et al., 2019) reinterpret a classifier $p_\theta(y|x)$ in supervised learning as an EBM for the joint distribution $p_\theta(x, y)$. Specifically, $p_\theta(y|x)$ is a categorical distribution:

$$p_\theta(y|x) = \frac{\exp(f_\theta(x)[y])}{\sum_y \exp(f_\theta(x)[y])}, \tag{2}$$

where $f_\theta(x) : \mathbb{R}^D \to \mathbb{R}^K$ maps each data point $x \in \mathbb{R}^D$ to K real-valued numbers known as logits, and $f_\theta(x)[y]$ indicates the logit of label y. JEM defines an EBM of the joint distribution with the same logits f_θ:

$$p_\theta(x, y) = \frac{\exp(f_\theta(x)[y])}{Z_\theta}, \tag{3}$$

where energy function $E_\theta(x, y) = -f_\theta(x)[y]$. To retain discriminative performance of $p_\theta(y|x)$, JEM factorizes the loglikelihood as

$$\log p_\theta(x, y) = \log p_\theta(x) + \log p_\theta(y|x), \tag{4}$$

and apply EBM training to benefit from generative models $p_\theta(x)$. Grathwohl et al. (2019) and Elflein et al. (2021) have shown that EBM training of the joint distribution improves calibration and out-of-distribution detection with various training methods.

Constrastive Learning. Our work is also related to contrastive learning, in that we approximate $\log p_\theta(x)$ and $\log p_\theta(x|y)$ by constrastive loss. Contrastive learning achieves remarkable success on downstream tasks, includes image classification, video understanding, knowledge distillation, etc (Khosla et al., 2020; Chen et al., 2020). In constrastive learning, a widely-used objective has the following form (Oord et al., 2018):

$$-\mathbb{E}_{p_{\text{data}}(x)}\left[\log \frac{\exp\{t_\theta(x)^\top t'_\theta(x)\}}{\sum_{i=1}^N \exp\{t_\theta(x)^\top t'_\theta(x_i)\}}\right], \tag{5}$$

where $t_\theta(x)$ and $t'_\theta(x)$ map each data point x to two different representation spaces. This is usually called InfoNCE loss. Different from existing methods for EBM training, we propose to leverage constrastive learning approximation for effective learning, without considering the generation ability such as SGLD, SSM and so on.

3 Triple-Hybrid Energy-based Model

Motivation. Many works (Grathwohl et al., 2019; Elflein et al., 2021; Du and Mordatch, 2019) have shown that EBMs could significantly reduce the expected calibration error and improve out-of-distribution detection for image classification. Specifically, the JEM proposed in Grathwohl et al. (2019) factorizes the joint distribution $\log p_\theta(x, y)$ into $\log p_\theta(x) + \log p_\theta(y|x)$, where $\log p_\theta(y|x)$ is to maintain the classification performance and $\log p_\theta(x)$ is the generative term which contributes to better calibration and out-of-distribution detection. On the contrary, the IGEBM proposed in Du and Mordatch (2019) factorizes the joint distribution $\log p_\theta(x, y)$ into $\log p_\theta(y) + \log p_\theta(x|y)$ for implicit generation and surprisingly find that it achieves better OOD performance. However, lack of $p_\theta(y|x)$ leads to terrible classification performance. It's shown in Grathwohl et al. (2019) that the classification accuracy dropped dramatically to 49.1% on the CIFAR10 dataset, while the accuracy is 92.9% by JEM.

On the other hand, Liu and Abbeel (2020) proposed a hybrid discriminative-generative energy-based model (HDGE) for both classification and generation. The loss function consists of a discriminative conditional log-likelihood $\log p_\theta(y|x)$ and a generative conditional log-likelihood $\log p_\theta(x|y)$. Compared to IGEBM, it includes $\log p_\theta(y|x)$ and thus achieves better classification performance. Compared to JEM, it includes the conditional generative model, rather than the marginal generative model. In other words, JEM targets to reduce the energy for data from the population $p_\theta(x)$, while HDGE aims at reducing the energy for compatible pair (x, y). This motivates us to combine the benefits of both conditional and marginal generative model for better calibration and OOD detection.

Triple-Hybrid Energy-based Model (THEM). We propose to make a hybrid model of the triple $\log p_\theta(y|x)$, $\log p_\theta(x|y)$ and $\log p_\theta(x)$, called Triple Hybrid Energy-based Model (THEM) and the objective function is

$$\mathbb{E}_{p_{\text{data}}(x,y)}\left[\log p_\theta(y|x) + \log p_\theta(x|y) + \log p_\theta(x)\right], \tag{6}$$

where $p_\theta(y|x)$ is the standard softmax neural classifer and the generative models $p_\theta(x), p_\theta(x|y)$ serve as regularization, always accompanied with better calibration and OOD detection.

From another perspective, we combine the two factorizations of the joint distribution $\log p_\theta(x, y)$ from JEM and IGEBM. We remark that the joint distribution can also be factorized as $(\log p_\theta(x) +$

$\log p_\theta(x|y) + \log p_\theta(y) + \log p_\theta(x|y))/2$. Our proposed THEM utilizes this factorization and treats $p_\theta(y)$ known as the label frequencies in data, which does not need to be optimized. Now we are ready to resolve the computational issues for THEM in the following.

Neural Classifier. The neural classifier term is easy to cope with. Specifically,

$$p_\theta(y|x) = \frac{\exp(f_\theta(x)[y])}{\sum_y \exp(f_\theta(x)[y])}, \tag{7}$$

where $f_\theta(x)[y]$ is the logit of label y. Thus we can derive the first term in (6) as the traditional cross-entropy loss:

$$\mathbb{E}_{p_{\text{data}}(x,y)} \left[\log p_\theta(y|x)\right]. \tag{8}$$

Conditional Generative Likelihood. The conditional generative likelihood can be derived from the joint distribution:

$$\begin{aligned} \log p_\theta(x|y) &= \log \frac{p_\theta(x,y)}{p_\theta(y)} = \log \frac{p_\theta(x,y)}{\sum_x p_\theta(x,y)} \\ &= \log \frac{\exp(f_\theta(x)[y])}{Z_\theta(y)}, \end{aligned} \tag{9}$$

where $Z_\theta(y) = \sum_x \exp(f_\theta(x)[y])$. By definition, this is also an EBM with energy function $E_\theta(x,y) = -f_\theta(x)[y]$. Energy-based models are well-known to be difficult to train. The Fenchel duality method used in Chen et al. (2021b) can estimate $Z_\theta(y)$. The stochastic gradient langevin dynamics (SGLD) adopted in Grathwohl et al. (2019) can approximate the gradient of $\log p_\theta(x|y)$. However, these methods require to calculate the derivative with respect to the input x and thus can't be applied to discrete data such as text tasks.

Approximation with Contrastive Learning. The above training methods are successful for generation purpose. Differently, we focus on classification with better calibration and OOD detection. As such, we propose to coarsely approximate the normalization constant as

$$Z_\theta(y) \approx \sum_{i=1}^{N} \exp(f_\theta(x_i)[y]), \tag{10}$$

where x_i is sampled from the data no matter whether y_i is equal to y or not. The second term in (6) is approximately

$$\mathbb{E}_{p_{\text{data}}(x,y)} \left[\log p_\theta(x|y)\right] \approx \log \frac{\exp(f_\theta(x)[y])}{\sum_{i=1}^{N} \exp(f_\theta(x_i)[y])}. \tag{11}$$

Since the samples for approximation are incorporated in the denominator using the same label y, the logits $f_\theta(x)[y]$ can be treated as the score function of input-label contrast (Rethmeier and Augenstein, 2021). As a result, this objective can be seen as the InfoNCE (5) in contrastive learning. Liu and Abbeel (2020) also proposed this approximation for image classification, while it's more suitable to text classification due to the discreteness of data.

To be more distinguishable between positive and negative samples but not concentrated on the nearest few samples (Zhang et al., 2021), we employ the temperature parameter τ in InfoNCE and the objective loss becomes

$$\log \frac{\exp(f_\theta(x)[y]/\tau)}{\sum_{i=1}^{N} \exp(f_\theta(x_i)[y]/\tau)}. \tag{12}$$

For the effectiveness of contrastive learning, it often requires a large number of negative samples (Chen et al., 2021a). Since directly increasing N is limited to hardware memory, we instead propose to use a memory bank (He et al., 2020) to store logits with negligible computational resources. In detail, we store the logits $f_\theta(x)[y]$ of the past samples into the memory bank.

Marginal Generative Likelihood. The marginal generative likelihood can be handled in the similar way as conditional generative likelihood. Specifically,

$$\begin{aligned} \log p_\theta(x) &= \log \left\{ \sum_y p_\theta(x,y) \right\} \\ &= \log \frac{\sum_y \exp(f_\theta(x)[y])}{Z_\theta}, \end{aligned} \tag{13}$$

where $Z_\theta = \sum_x \sum_y \exp(f_\theta(x)[y])$. We propose to approximate the third term in (6) as

$$\mathbb{E}_{p_{\text{data}}(x,y)} \left[\log p_\theta(x)\right] \approx \log \frac{\sum_y \exp(f_\theta(x)[y])}{\sum_{i=1}^{N} \sum_y \exp(f_\theta(x_i)[y])}, \tag{14}$$

where x_i is sampled from the data distribution. As the conditional generative likelihood, techniques of temperature parameter and memory bank are also employed.

4 Experiments

In this section, we conduct thorough experiments to investigate the empirical peformance of our proposed methods. We first introduce the criteria for ID calibration and OOD detection.

ID Calibration. For a well-calibrated model, the confidence estimate $\hat{p}$ of the model is expected to be comparable to true probability (accuracy): $\mathbb{P}(\hat{y} = y|\hat{p}) = \hat{p}$ (Desai and Durrett, 2020; Kong et al., 2020). The calibration error for a given confidence $p \in (0, 1)$ is defined as the followings:

$$\mathbb{E}_p = |\mathbb{P}(\hat{y}(x) = y(x)|\hat{P}(x) = p) - p|, \quad (15)$$

where $\hat{y}(x)$ is the label predicted by the model, $y(x)$ is the true label for input x and $\hat{P}(x)$ is the output probability associated with the predicted label $\hat{y}(x)$. To evaluate the overall calibration error, we partition $(0, 1)$ into M bins of equal size and let b_m denote the set of prediction confidences which lie in the m-th bin. The expected calibration error (ECE) is calculated by weighting the difference between accuracy and confidence of each bin:

$$\begin{aligned} \text{acc}(b_m) &= \frac{1}{|b_m|} \sum_{i \in b_m} \mathbb{I}(\hat{y}_i = y). \\ \text{conf}(b_m) &= \frac{1}{|b_m|} \sum_{i \in b_m} \hat{p}_i. \\ \text{ECE} &= \sum_{m=1}^{M} \frac{|b_m|}{N} |\text{acc}(b_m) - \text{conf}(b_m)|. \end{aligned} \quad (16)$$

OOD Detection. In general, OOD detection is a binary classification problem, where the model is required to produce a score $s_\theta(x) \in \mathbb{R}$. Usually we can set a threshold δ to detect OOD samples whose score functions are below the threshold. A well-calibrated model is expected to output higher scores for in-distribution examples than out-of-distribution examples. A widely used score function is maximum prediction probability (Hendrycks and Gimpel, 2016):

$$s_\theta(x) = \max_y p_\theta(y|x). \quad (17)$$

Following Kong et al. (2020), we employ the empirical Normalized Bounded Area Under the Calibration Curve (NBAUCC) as the evaluation metric rather than the Area Under the Receiver-Operating curve (AUROC; Hendrycks and Gimpel, 2016) and the Area Under the Precision-Recall curve (AUPR; Elflein et al., 2021). The main reason is that we would like to use a threshold as low as possible to detect ODD samples and more details are referred to Kong et al. (2020).

Target. In our experiments, we are interested in answering the following questions:

1. Does THEM achieve better calibration compared to baselines?
2. Does THEM improve OOD detection?
3. The effect of temperature and the size of memory bank on THEM, JEM(CL) and HDGE(CL).

Datasets. We consider the eight datasets of GLUE used in He et al. (2021) to evaluate the ID calibration, since there are no out-of-distribution samples in GLUE. We use the official code[1] to acquire the development and test dataset of GLUE. Furthermore, we consider six more datasets used in Kong et al. (2020) to evaluate both ID calibration and OOD detection. Details of the datasets are in Table 4 and 5 in Appendix.

Baselines. For GLUE datasets, we compare our method against that of He et al. (2021), which is state-of-the-art EBMs on natural language understanding models. Their method is based on Residual-EBM which can work with more flexible energy functions, but the computational cost is also huge compared to our method. We also compare with other three strong baselines for calibration: finetune, Scal-bin and T-scale used in He et al. (2021). For fair comparisons, we follow the experiment settings in their work. We use Roberta as the backbone and the bins of ECE is set to 20.

For the additional six datasets, we compare our methods with nine strong baselines in Kong et al. (2020) including (1) BERT finetuning, (2) Post-calibration method: Temperature Scaling (TS; Guo et al., 2017), (3) Model ensemble: Monte Carlo Dropout (MCDP; Gal and Ghahramani, 2016), (4) Over-confident correction: Label Smoothing (LS; Müller et al., 2019), Entropy Regularized Loss (ERL; Pereyra et al., 2017), Virtual Adversarial Training (VAT; Miyato et al., 2018), and (5) Data-augmentation: Mixup (Zhang et al., 2018), Manifold-Mixup (M-Mixup; Verma et al., 2019), and Manifold-regularization (M-regularization; Kong et al., 2020). We use BERT as the backbone and the bins of ECE is set 15 just as Kong et al. (2020). Besides, we also use NBAUCC as the misclassification evaluation to make fair and

[1]We use the official code: https://github.com/salesforce/ebm_calibration_nlu

comprehensive comparisons. For these datasets, we don't compare our method with that of He et al. (2021) since it is time-consuming and needs more computational resources to finetune a noise distribution and generate negative samples for NCE training. At last, we use $\text{NBAUCC}_{0.5}$ as the evaluation metric for OOD detection. The OOD datasets often need an ID dataset for training and an OOD dataset for OOD detection evaluation. More details of datasets can be found in Appendix A.

Implementation Details. We employ ADAM (Kingma and Ba, 2014) as the optimizer for all experiments with the following parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, gradient clip of 1.0, and L_2 weight decay of 0.1. We search learning rate in $[1e^{-5}, 2e^{-5}, 3e^{-5}, 5e^{-5}]$ with the training epochs in $[2, 3, 5, 10]$. Our model is built with a classifier on the top of the pretrained language models including BERT (Devlin et al., 2019) and Roberta (Liu et al., 2019) using the implementation of Huggingface (Wolf et al., 2020). For contrastive learning, we set the size of memory bank N to 65536 and the temperature τ to 0.1. All experiments run 5-times and we report the average performance on test dataset. The test result is selected based on loss and accuracy on the development dataset[2]. All experiments are conducted on a single NVIDIA RTX 2080TI 12G GPU. Our implementation is based on the official codes of MoCo[3] and Manifold-regularization[4].

Results on ID Calibration. Table 1 and 2 show the accuracy (acc) and ECE results for GLUE tasks and the six additional datasets respectively, with different baseline methods. Except our proposed THEM, we also include (1) HDGE: $\log p(x|y) + \log p(y|x)$, (2) JEM: $\log p(y|x) + \log p(x)$, (3) IGEMB: $\log p(x|y)$ but trained with contrastive learning (CL) proposed in this paper. These three EBMs are trained by MCMC for images in previous literatures, while we are the first to train them by contrastive learning for NLU tasks.

From Table 1, EBMs with contrastive learning achieves significant improvements on ECE with competitive accuracy, compared to He et al. (2021). He et al. (2021) redefines energy function based on Residual-EBM and estimates parameters using NCE with a finetuned GPT (Radford et al.) on dataset as noise distribution. Not only does it need more computing resources to finetune a GPT model for each dataset, but also the quantity and quality of negative samples generated by noise distribution have big impacts for accurate parameters estimation using NCE (He et al., 2021; Gutmann and Hyvärinen, 2010). In contrast, our model achieves better results with negligible computational resource compared to standard finetuning.

From Table 2, our method achieves the best ECE on six multiclass datasets on various domains. Compared to M-regularization (Kong et al., 2020) which is specifically designed to prevent overconfident predictions for both in-distribution and out-of-distribution, our framework without an explicit calibration mechanism achieves the best ECE, demonstrating the effectiveness of EBMs trained with contrastive learning paradigm. On average, the result of the proposed THEM is very close to the best one in terms of ECE.

Results on OOD Detection. In general, OOD detection is a binary classification problem, where the model is required to produce a score for a data point to detect whether it is an ID or OOD sample. Here we use equation (17) as score function and $\text{NBAUCC}_{0.5}$ as evaluation metric. Table 3 summarizes the $\text{NBAUCC}_{0.5}$ for misclassification detection and OOD detection. It can be seen that compared with all baselines, especially the strong baseline M-regularization (Kong et al., 2020), our method achieves the best misclassification on all data sets with significant improvements. In terms of OOD detection, our results averaged on six datasets are comparable to the performance of M-regularization and are superior to other baselines on all datasets except M-regularization on Yahoo. These results shows that THEM provides a simple yet effective way to improve OOD detection.

Analysis of Generative Density. From Table 1 and Table 2, the generative density including marginal data density $p_\theta(x)$ and class conditional data density $p_\theta(x|y)$ are mainly contributed to the improvements of ID calibration and OOD detection compared to standard finetuning and previous calibration methods. However, different generative terms may have different impacts on final performance. In most NLU tasks, JEM(CL) achieves better ID and OOD calibration performance compared to HDGE(CL) which is different from the experimental results observed in Liu and Abbeel (2020) on

[2]Following He et al. (2021), we don't use ECE as the metric to select the best model for evaluation

[3]https://github.com/facebookresearch/moco

[4]https://github.com/Lingkai-Kong/Calibrated-BERT-Fine-Tuning

Table 1: Test-set accuracy and ECE results for different methods on GLUE tasks. The leading zeros are omitted to save space. Note that the hyperparameters of T-Scal and Scal-bin are searched on the development dataset and applied to test dataset. The average value is compute on all nine test sets. For each task, the method that achieves best calibration are shown in bold.

Method	SST-2		MNLI		MNLI(mm)		QNLI		QQP		MRPC		COLA		RTE		WNLI		AVG	
	acc	**ECE**	**acc**	**ECE**	**acc**	**ECE**	**acc**	**ECE**	**acc**	**ECE**	**acc**	**ECE**	**acc**	**ECE**	**acc**	**ECE**	**acc**	**ECE**	**acc**	**ECE**
Baseline (He et al., 2021)																				
finetune	.942	.050	.876	.067	.872	.068	.929	.043	.904	.034	.862	.133	.539	.182	.724	.279	.571	.058	.802	.102
Scal-bin(dev)	.944	**.019**	.876	.030	.870	.032	.931	.021	.905	.021	.862	.062	.557	**.048**	.731	**.042**	.542	.189	.802	.052
T-Scale(dev)	.942	.037	.876	.024	.872	.026	.929	.018	.904	.026	.862	.126	.539	.109	.724	.235	.571	.046	.802	.072
Residual-EBM-NCE (He et al., 2021)																				
ebm-scalar	.942	.033	.871	.038	.871	.047	.927	**.016**	.899	.034	.862	.098	.540	.150	.753	.207	.542	**.033**	.801	.073
ebm-hidden	.956	.032	.869	.032	.868	.044	.923	**.016**	.900	.033	.867	.099	.545	.131	.797	.148	.542	.036	.807	.063
ebm-s-hidden	.947	.038	.875	**.027**	.872	.031	.930	**.016**	.900	.032	.862	.089	.563	.133	.811	.182	.571	.073	.815	.069
Ours																				
HDGE(CL)	.938	.036	.870	.040	.864	.049	.927	.024	.908	.023	.862	.056	.539	.101	.753	.069	.571	.051	.803	.048
JEM(CL)	.926	.043	.872	.033	.868	**.023**	.927	.021	.907	**.009**	.877	.060	.562	.107	.753	.073	.571	.057	.806	**.047**
IGEBM(CL)	.922	.054	.868	.124	.869	.125	.931	.065	.910	.087	.867	**.029**	.549	.060	.789	.052	.571	.044	.808	.071
THEM	.922	.035	.867	.043	.866	.043	.928	.028	.910	.019	.872	.062	.551	.085	.724	.082	.571	.050	.801	**.049**

Table 2: ECE and accuracy (in percentage) on test set for different methods on six multiclass datasets listed in Table 5. We report the average performance of 5 random initializations. For each task, the method that achieves best calibration are shown in bold.

Method	20NG_{15}		20NG		WOS_{100}		WOS		Yahoo_{8}		Yahoo		AVG	
	acc	**ECE**	**acc**	**ECE**	**acc**	**ECE**	**acc**	**ECE**	**acc**	**ECE**	**acc**	**ECE**	**acc**	**ECE**
baseline (Kong et al., 2020)														
BERT	87.42	9.24	84.55	11.61	81.94	6.81	79.40	6.74	73.58	10.11	71.89	10.54	79.79	9.17
TS	87.42	4.42	84.55	8.17	81.94	3.63	79.40	4.43	73.58	5.18	71.89	4.24	79.79	5.01
MCDP	87.45	6.88	84.55	9.17	82.09	4.00	79.67	3.55	73.67	6.54	71.99	6.72	79.90	6.14
LS	87.54	4.35	85.02	6.15	81.95	4.35	79.47	4.67	73.66	4.89	71.54	3.61	79.86	4.67
ERL	87.67	7.16	84.83	6.10	81.96	3.74	79.48	3.35	73.63	3.42	72.01	2.96	79.92	4.45
VAT	87.61	9.07	85.20	11.28	81.65	7.27	79.71	6.76	73.71	10.96	72.08	7.92	79.99	8.87
Mixup	87.49	5.98	84.86	9.02	81.97	4.72	79.51	4.21	73.88	4.60	71.82	5.18	79.92	5.61
M-Mixup	87.40	5.04	84.45	7.78	81.77	6.48	79.57	6.68	72.03	7.01	72.03	6.07	79.54	6.51
M-regularization	87.44	3.69	84.53	4.43	81.59	3.24	79.06	**3.04**	73.71	3.03	72.17	3.42	79.75	3.47
Ours														
HDGE(CL)	87.34	4.71	84.47	7.76	81.14	4.00	78.68	4.12	73.53	4.02	71.62	5.97	79.46	5.09
JEM(CL)	87.98	3.10	84.81	**2.17**	81.80	3.47	78.74	3.27	73.72	2.17	72.60	**1.64**	79.86	**2.58**
IGEBM(CL)	88.35	2.47	84.06	3.87	81.51	11.72	78.46	13.73	73.01	4.00	70.82	2.01	79.36	6.30
THEM	88.35	**2.09**	84.99	3.91	81.05	**3.01**	78.72	3.19	73.80	**1.55**	72.00	2.19	79.81	**2.65**

computer vision tasks. The main reason may lie on the estimation methods that it is more stable and effective to approximate the log-likelihood with contrastive loss compared to MCMC or score-matching for calibration, when the generation ability is not under consideration. While our model achieves comparable ID calibration performance across various datasets on average and better OOD detection.

The hyper-parameters study of InfoNCE on ID calibration and OOD detection. To study the effect of the hyper-parameters including the sample size, and temperature of InfoNCE for training THEM, JEM(CL) and HDGE(CL), we conduct numerous experiments on **20NG**, **Yahoo** and **WOS** dataset. Due to the limited computational resources, we set temperature to 0.1 and vary the size of memory bank from 128 to 65536 to study the effect of the memory bank size on ECE and OOD. Similarly, we set the memory bank size to 65536 and vary the temperature from 0.01 to 0.2 to study the effect of the temperature on ECE and OOD.

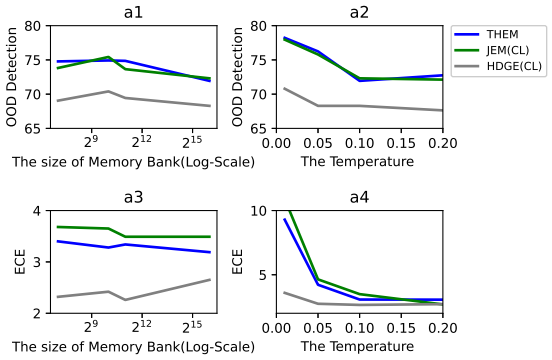

Figure 1: The effect of the memory bank size and temperature for ECE and OOD on **WOS** dataset.

Table 3: NBAUCC$_{0.5}$ on misclassification detection and OOD detection (in percentage) for different methods on six multiclass datasets listed in Table 5. We report the average performance of 5 random initializations.

Data (OOD)	Misclassification(↑)							OOD Detection(↑)						
	20NG$_{15}$	20NG	WOS$_{100}$	WOS	Yahoo$_8$	Yahoo	AVG	20NG$_{15}$ 20NG$_5$	20NG SST-2	WOS$_{100}$ WOS$_{34}$	WOS AGnews	Yhaoo$_8$ Yhaoo$_2$	Yahoo Yelp	AVG
baseline (Kong et al., 2020)														
BERT	2.30	2.86	16.53	20.52	7.47	8.43	9.68	2.66	21.65	23.12	49.84	8.35	13.88	19.91
TS	6.08	5.74	21.20	23.76	10.48	12.74	13.33	6.62	32.64	28.12	53.32	11.55	20.27	25.42
MCDP	4.37	5.28	20.44	24.16	10.12	10.75	12.52	3.99	25.10	27.28	53.52	9.98	15.93	22.63
LS	4.72	6.75	20.37	23.56	11.19	16.15	13.79	5.70	41.08	27.12	58.48	12.02	19.81	27.36
ERL	8.54	10.35	20.49	25.13	12.89	15.47	15.47	8.78	47.00	27.73	56.67	13.78	23.47	29.57
VAT	2.52	3.36	18.70	19.96	6.54	10.37	10.24	2.96	29.62	23.41	54.60	7.42	17.65	22.61
Mixup	4.99	4.51	20.65	24.80	10.75	11.29	12.83	5.86	31.84	26.77	58.02	11.62	19.84	25.65
M-mixup	2.16	3.16	16.94	19.39	9.09	11.79	10.42	2.36	26.08	24.08	51.39	10.08	22.41	22.73
M-regularization	9.10	10.76	26.93	30.80	14.34	17.88	18.30	9.69	63.92	35.60	71.13	14.94	**29.40**	**37.44**
Ours														
HDGE(CL)	7.99	6.68	25.25	27.82	12.31	14.72	15.79	7.42	57.09	34.81	68.29	11.55	20.62	33.29
JEM(CL)	**15.31**	14.88	25.55	32.97	**16.25**	16.16	20.18	12.23	61.99	34.70	**72.31**	16.17	19.80	36.20
IGEBM(CL)	13.87	**15.34**	14.37	15.64	14.52	**21.83**	15.92	**14.47**	**64.75**	23.67	57.94	**17.93**	24.22	33.83
THEM	11.56	11.11	**31.82**	**33.02**	16.11	18.25	**20.31**	9.36	62.86	**40.16**	71.94	17.28	19.73	**36.88**

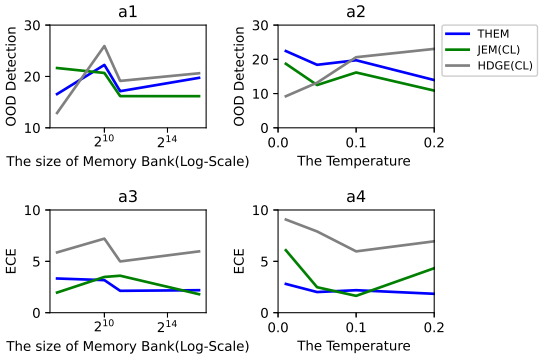

Figure 2: The effect of the memory bank size and temperature for ECE and OOD on **Yahoo** dataset.

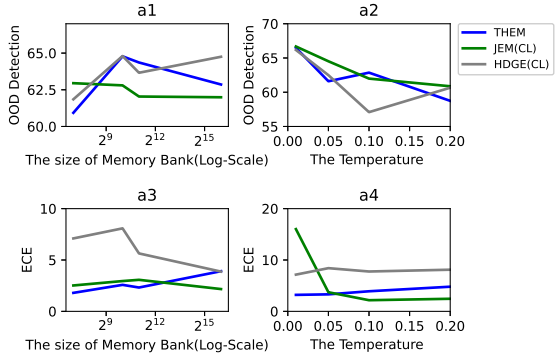

Figure 3: The effect of the memory bank size and temperature for ECE and OOD on **20NG** dataset.

From the trends of memory bank size and temperature on WOS(134-class) in Fig 1, HDGE performs better than THEM and JEM(CL) in terms of ECE. However, it performs significantly worse than THEM and JEM(CL) in terms of OOD. On the other hand, they are all stable in terms of ECE and OOD from the trend of memory bank size.

However, for Yahoo(10-class) in Fig 2 and 20NG(20-class) in Fig 3, from the trend of memory bank size, THEM and HDGE are superior to JEM(CL) in terms of OOD. THEM and JEM(CL) perform better than HDGE in terms of ECE. From the trend of temperature, THEM performs better than JEM(CL) and HDGE in all evaluation metrics. In general, THEM is more stable in terms of ECE from the trend of temperature and memory bank size.

5 Conclusion

In our work, we propose a triple-hybrid EBM with combination of classifier, conditional generative model and marginal generative model into a unified framework called THEM. To train EBMs effectively and efficiently, we leaverage contrastive learning to approximate the log-likelihood of EBMs with negligible computational resources. Extensive experiments demonstrates that our model outperforms the state-of-art methods in terms of ID calibration and OOD detection with competitive accuracy. We further apply contrastive learning to JEM and IGEBM without considering the generation ability to obtain JEM(CL) and IGEBM(CL) respectively. Compared to JEM(CL) and HDGE(CL), our model is more robust to the hyper-parameters of contrastive learning including the temperature and size of memory bank in terms of ID calibration and OOD detection.

6 Limitations

In our work, our model is derived from the perspective of EBMs. However, it lacks of generation ability due to the approximation of log-likelihoods with contrastrive learning which may limit the

power of generative modeling such as data augmentation (Grathwohl et al., 2019). We will explore MCMC-based methods such as (Eikema et al., 2021; Qin et al., 2022) to train THEM to take advantage of generative modeling. As for OOD detection, we only use Maximum Prediction Probability. But many other OOD scoring functions are proposed from the perpective of EBMs (Ouyang et al., 2021; Zhou et al., 2021; Liu et al., 2020; Elflein et al., 2021; Grathwohl et al., 2019). And it may be explored in future works to study the OOD performance with different scoring functions.

7 Acknowledgements

This research of Zhang is supported by the National Natural Science Foundation of China (NSFC grant nos. 12101241).

References

Pavel Blinov, Manvel Avetisian, Vladimir Kokh, Dmitry Umerenkov, and Alexander Tuzhilin. 2020. Predicting clinical diagnosis from patients electronic health records using bert-based neural networks. In *International Conference on Artificial Intelligence in Medicine*, pages 111–121. Springer.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Aaai*, volume 2, pages 830–835.

Junya Chen, Zhe Gan, Xuan Li, Qing Guo, Liqun Chen, Shuyang Gao, Tagyoung Chung, Yi Xu, Belinda Zeng, Wenlian Lu, et al. 2021a. Simpler, faster, stronger: Breaking the log-k curse on contrastive learners with flatnce. *arXiv preprint arXiv:2107.01152*.

Si-An Chen, Chun-Liang Li, and Hsuan-Tien Lin. 2021b. A unified view of cgans with and without classifiers. *Advances in Neural Information Processing Systems*, 34.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. 2019. Residual energy-based models for text generation. In *International Conference on Learning Representations*.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302.

J. Devlin, M. W. Chang, K. Lee, and K Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yilun Du and Igor Mordatch. 2019. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32.

David Duvenaud, Jacob Kelly, Kevin Swersky, Milad Hashemi, Mohammad Norouzi, and Will Grathwohl. 2021. No mcmc for me: Amortized samplers for fast and stable training of energy-based models.

B. Eikema, G. Kruszewski, H. Elsahar, and M. Dymetman. 2021. Sampling from discrete energy-based models with quality/efficiency trade-offs.

Sven Elflein, Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. 2021. On out-of-distribution detection with energy-based models. *arXiv preprint arXiv:2107.08785*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. 2019. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Tianxing He, Bryan McCann, Caiming Xiong, and Ehsan Hosseini-Asl. 2021. Joint energy-based model training for better calibrated natural language understanding models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1754–1761.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Taehee Jung, Dongyeop Kang, Hua Cheng, Lucas Mentch, and Thomas Schaaf. 2020. Posterior calibrated training on sentence classification tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2723–2730.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Shivesh Khaitan, Qin Lin, and John M Dolan. 2021. Safe planning and control under uncertainty for self-driving. *IEEE Transactions on Vehicular Technology*, 70(10):9826–9837.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in-and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340.

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 364–371. IEEE.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, Hong Yu, et al. 2019. Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: an empirical study. *JMIR medical informatics*, 7(3):e14830.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. A survey of pretrained language models based text generation. *arXiv preprint arXiv:2201.05273*.

Linfeng Li, Peng Wang, Jun Yan, Yao Wang, Simin Li, Jinpeng Jiang, Zhe Sun, Buzhou Tang, Tsung-Hui Chang, Shenghui Wang, et al. 2020. Real-world data medical knowledge graph: construction and applications. *Artificial intelligence in medicine*, 103:101817.

Hao Liu and Pieter Abbeel. 2020. Hybrid discriminative-generative training via contrastive learning. *arXiv preprint arXiv:2007.09070*.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Yawen Ouyang, Jiasheng Ye, Yu Chen, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2021. Energy-based unknown intent detection with data manipulation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2852–2861.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

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

L. Qin, S. Welleck, D. Khashabi, and Y. Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Amir Rahimi, Amirreza Shaban, Ching-An Cheng, Richard Hartley, and Byron Boots. 2020. Intra order-preserving functions for calibration of multi-class neural networks. *Advances in Neural Information Processing Systems*, 33:13456–13467.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

Nils Rethmeier and Isabelle Augenstein. 2021. A primer on contrastive pretraining in language processing: Methods, lessons learned and perspectives. *arXiv preprint arXiv:2102.12982*.

Sarah Sarabadani. 2019. Detection of adverse drug reaction mentions in tweets using elmo. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 120–122.

Richard Socher, Yoshua Bengio, and Christopher D Manning. 2012. Deep learning for nlp (without magic). In *Tutorial Abstracts of ACL 2012*, pages 5–5.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. 2020. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pages 574–584. PMLR.

Ting Sun and Miklos A Vasarhalyi. 2021. Predicting credit card delinquencies: An application of deep neural networks. In *Handbook of Financial Econometrics, Mathematics, Statistics, and Machine Learning*, pages 4349–4381. World Scientific.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079.

Max Welling and Yee W Teh. 2011. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Oliver Zhang, Mike Wu, Jasmine Bayrooti, and Noah Goodman. 2021. Temperature as uncertainty in contrastive learning. *arXiv preprint arXiv:2110.04403*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021. Contrastive out-of-distribution detection for pretrained transformers. *arXiv preprint arXiv:2104.08812*.

A Dataset

The details of dataset for evaluation of in-distribution ECE and out-of-distribution detection.

Table 4: The detail information about GLUE

dataset	task	labels	train/dev/test
RTE	Similarity	2	2.5k/0.14k/0.14k
CoLA	Grammatical	2	8.5k/0.51k/0.51k
WNLI	Entailment	2	3.1k/0.03k/0.03k
MRPC	Paraphrase	2	3.7k/0.20k/0.20k
QNLI	Entailment	2	108k/2.5k/2.5k
MNLI-m	Entailment	3	393k/4.8k/4.8k
MNLI-mm	Entailment	3	393k/4.4k/4.4k
QQP	Paraphrase	2	364k/20k/20k
SST-2	Classification	2	67k/0.43k/0.43k

Table 5: The detail information about six multiclass-datasets

in-distribution out-of-distribution	labels	train/dev/test
$20NG_{15}$ $20NG_5$	15 5	7k/1.7k/5.8k -/-/1.7k
20NG SST-2	20 2	9k /2.2k/7.5k -/-/1.8k
WOS_{100} WOS_{34}	100 34	16k/4.1k/14k -/-/ 4.8k
WOS AGnews	134 4	22k/5.6k /18k - /-/ 7.6k
$Yahoo_8$ $Yahoo_2$	8 2	16k/4k/48k -/-/12k
Yahoo Yelp	10 2	20k/5k/60k -/-/38k

1. 20NG[5]. The 20 Newsgroups dataset (20NG) contains news articles with 20 categories. We use Stanford Sentiment Treebank (SST-2) (Socher et al., 2012) as the OOD data.
2. $20NG_{15}$. We take the first 15 categories of 20NG as the in-distribution data and the other 5 categories ($20NG_5$) as the OOD data.

[5]We use the 20 Newsgroups dataset from: http://qwone.com/~jason/20Newsgroups/

3. WOS (Kowsari et al., 2017). Web of Science (WOS) dataset contains 134 categories of scientific articles. We use AGnews (Zhang et al., 2015) as the OOD data.
4. WOS_{100}. We use the first 100 classes of WOS as the in-distribution data and the other 34 classes (WOS_{34}) as the OOD data.
5. Yahoo (Chang et al., 2008). This dataset contains 10 categories posted to 'Yahoo!Answers' of questions. We randomly draw 2000 from 140,000 samples for each category as the training set. We use Yelp (Zhang et al., 2015) as the OOD data.
6. Yahoo_8. We use the first 8 classes of Yahoo as the in-distribution data and the other 2 classes (Yahoo_2) as the OOD data.