The Open-World Lottery Ticket Hypothesis for OOD Intent Classification

Yunhua Zhou[1,2]*, **Pengyu Wang**[1]*, **Peiju Liu**[1], **Yuxin Wang**[1], **Xipeng Qiu**[1†]
[1]School of Computer Science, Fudan University [2]Shanghai AI Laboratory
zhouyunhua@pjlab.org.cn xpqiu@fudan.edu.cn
{pywang22, pjliu21, wangyuxin21}@m.fudan.edu.cn

Abstract

Most existing methods of Out-of-Domain (OOD) intent classification rely on extensive auxiliary OOD corpora or specific training paradigms. However, they are underdeveloped in the underlying principle that the models should have differentiated confidence in In- and Out-of-domain intent. In this work, we shed light on the fundamental cause of model overconfidence on OOD and demonstrate that calibrated subnetworks can be uncovered by pruning the overparameterized model. Calibrated confidence provided by the subnetwork can better distinguish In- and Out-of-domain, which can be a benefit for almost all *post hoc* methods. In addition to bringing fundamental insights, we also extend the Lottery Ticket Hypothesis to open-world scenarios. We conduct extensive experiments on four real-world datasets to demonstrate our approach can establish consistent improvements compared with a suite of competitive baselines.

Keywords: Lottery Ticket, OOD Intent Classification

1. Introduction

Interactive Systems, such as Task-Oriented Dialog Systems(TODS), are gradually integrating into and facilitating the daily life of people. However, in open-world scenarios, i.e., the training and test set come from the different distributions or domains, it is often encountered that the expressed intents are reasonable but beyond the domains supported by the Interactive Systems, resulting in mapping the intent to the wrong subsequent processing pipelines. Therefore, Interactive Systems not only need to maintain performance in In-Domain (IND) intents but also need to correctly identify Out-of-Domain (OOD) intents.

Recently, to get critical insights into *Does the model knows what it does not know?* i.e., the model should be high-confident in IND and low-confident in OOD (due to unseen), Hendrycks et al. (2020) take a step to show that compared with previous models, such as LSTMs, the confidence scores produced by the Pre-Trained Models'maximum softmax probabilities can significantly distinguish IND and OOD but remain a long way before it is perfect.

What prevents the confidence of the model from being further trusted? Current efforts have primarily concentrated on developing appropriate *post hoc* (i.e., not involved in training and after training) methods or scoring functions based on maximum softmax probability, such as MSP (Hendrycks et al., 2020), Entropy (Liu et al., 2020), to measure OOD uncertainty. Despite the advancements made, these approaches are inherently limited and lack broad applicability, as the root underlying causes have yet to be thoroughly investigated and understood.

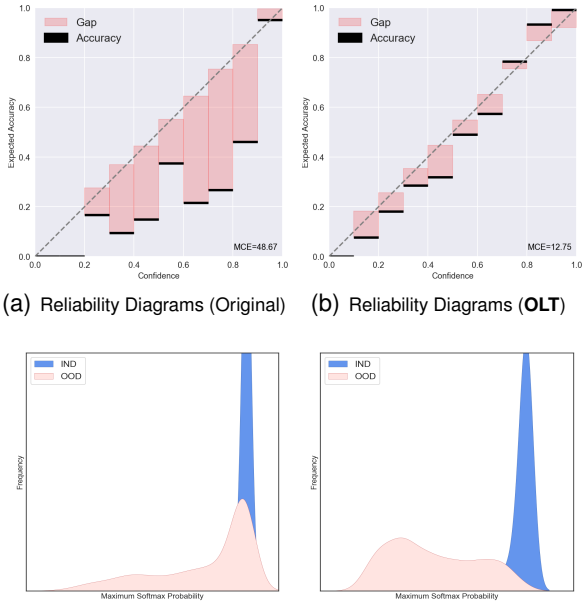
We take a step forward and observe that the maximum softmax probability outputted by the overparameterized model cannot correctly reflect the confidence of the model, which is known as poor-calibrated (Guo et al., 2017) and can be visualized [1] by **reliability diagrams** as shown in Fig.1(a). When encountering open-world scenarios, the unreliable predicted confidence (and other *post hoc* measures based on it) given by the poor-calibrated model cannot be measured to uncertain about samples correctly. Furthermore, subsequent analysis (Section 3.2) shows that the overparameterized model tends to be overconfident, which is also consistent with the experiment as shown in Fig.1(c). This phenomenon undermines the underlying idea that the model should be much low-confident in OOD and makes it non-trivial to distinguish between IND and OOD.

In this paper, in addition to giving fundamental insight, we also explore how to calibrate the model to provide reliable confidence. To this end, we first set out to establish the effect of overparameterization in poor calibration and theoretically demonstrate overparameterization would aggravate overconfident predictions on OOD inputs. Inspired by this, different from the previous work, we do not design a specific simple method to measure OOD uncertainty. Instead, through masking the parameters that are not of interest to the target task, we prune a calibrated subnetwork from an overparameterized Pre-Trained model during training, which has more general reliable confidence to better differen-

*Equal contribution.
†Corresponding author.

[1]https://github.com/hollance/reliability-diagrams

LREC-COLING 2024, pages 15988–15999
20-25 May, 2024.



(a) Reliability Diagrams (Original) (b) Reliability Diagrams (**OLT**)

(c) Distribution of Scores (Original) (d) Distribution of Scores (**OLT**)

Figure 1: Plots showing (**Top**) Reliability diagrams and (**Bottom**) The distribution of In-and Out-of-domain uncertainty scores in the Stackoverflow dataset. The OLT denotes our proposed Open-world Lottery Ticket. The reliability diagrams (pink) are about the function of confidence, which measures the gap (i.e., miscalibration) between expected sample accuracy (black) and confidence. The Maximum Calibration Error (MCE) measures the maximum gap. If a model meets perfect calibration, the gap is zero and the diagrams disappear.

tiate IND and OOD and can be a benefit for almost all *post hoc* methods.

Especially, beyond the established awareness that temperature scaling can help improve calibration in the *post hoc* phase empirically Guo et al. (2017), we contribute a new general insight on temperature scaling in open-world scenarios and theoretically demonstrate temperature scaling can substantially differentiate IND and OOD.

Going further, combined with the above calibration of subnetwork and *post hoc* measure (temperature scaling adopted in this paper), we can further generalize the Lottery Ticket Hypothesis (Frankle and Carbin, 2019) to the open-world. The **O**pen-world **L**ottery **T**icket **H**ypothesis (OLTH) is articulated as:

An initialized overparameterized neural network contains a winning subnetwork—through one-shot pruning and minor post-processing, which can match the commensurate performance in IND identification as original, but also better detect OOD at a commensurate training cost as the original.

Compared with the original Lottery Ticket Hypothesis, we generate **O**pen-world **L**ottery **T**icket (OLT) through one-shot pruning without iterative pruning. The OLT could be better-calibrated, as shown in Fig.1(b) and not only guarantees the precision of IND recognition but can better distinguish between IND and OOD, as shown in Fig.1(d), signifying its adaptability to the open-world. Extensive experiments are conducted on four real-world datasets and further verify our hypothesis. Our contributions and insights are as:

(Theory) We establish the effect of overparameterization in overconfidence and demonstrate that the well-calibrated confidence of the subnetwork can help improve OOD detection. Furthermore, we empirically extend the LTH–we can identify a lottery ticket from the overparameterized model that is more suitable for the open-world setting.

(Methodology) We propose a one-shot (without Iterative) Magnitude Pruning to uncover the lottery ticket of interest to the target task, which has more general reliable confidence to better differentiate IND and OOD and can be a benefit for almost all *post hoc* uncertainty measurements.

(Experiments) Extensive experiments and analysis show that our method can improve OOD detection on the premise of the accuracy of IND recognition, which confirms the correctness of the Open-world Lottery Ticket Hypothesis.[2]

2. Related Work

There are two types of work close to our research–Out-of-domain Detection and Sparse Network.

Out-of-domain Detection This kind of research mainly focuses on how to design appropriate scoring functions to detect OOD. Hendrycks and Gimpel (2017) adopt the maximum softmax probability (**MSP**) and provide several baselines for the follow-up research. Liang et al. (2018) (**ODIN**) add small perturbations to inputs and temperature to softmax score based on maximum softmax probability. Lee et al. (2018) detect OOD samples by calculating the **Mahalanobis distance** between the sample and the different In-domain distributions. Zheng et al. (2020) distinguish IND and OOD by **Entropy** calculated on softmax probability. Liu et al. (2020) regard the **Energy** score calculated from output logits as a better scoring function. Sun et al. (2021) propose a simple *post hoc* OOD detection method by rectifying the activations (**ReAct**) output in the penultimate layer of model. Hendrycks et al. (2019) propose that under large-scale and real-world settings, taking **MaxLogit** as the scoring function is better than maximum softmax probability.

[2]Codes is publicly available at: https://github.com/zyh190507/Open-world-Lottery

Sparse Network Our approach is also inspired by the work related to sparse networks. Louizos et al. (2017) prune the network by adding $\mathcal{L}_0$ norm regularization on parameters. Frankle and Carbin (2019) propose the Lottery Ticket Hypothesis–a subnetwork (winning ticket) with comparable performance as the original network can be uncovered from the randomly initialized overparameterized network at the same (or no more than) cost of the training original network. Zhang et al. (2021a); Zheng et al. (2022) propose that the structure of the model is related to the spurious correlation and an unbiased substructure can be found from the biased model. Based on Louizos et al. (2017), Cao et al. (2021a) can search for various subnetworks that perform the various linguistic tasks of interest.

3. Proposed Method

3.1. Problem Statement

OOD intent classification usually adheres to the following paradigm: Denote $Y := \{1,\ldots,k\}$ as the pre-defined intent set in the TODS where k is the number of intents and $\mathbb{X}$ as the whole input space. For an utterance $x \in \mathbb{X}$, the logits about intents can be output through a neural network $\mathcal{F} : \mathbb{X} \to R^{[Y]}$. A desirable scoring function (also known as decision function) $\mathcal{G}$, which can detect OOD intent while ensuring the accuracy of the identification of known intents, is the objective of OOD intent classification. The prediction can be formed as:

$$\hat{Y} = \begin{cases} \text{OOD}, & \mathcal{G}(x,\mathcal{F}) < \theta, \\ \text{argmax}_{k\in[Y]}\phi_k(\mathcal{F}(x)), & \mathcal{G}(x,\mathcal{F}) \geq \theta. \end{cases}, \quad (1)$$

where ϕ is a function of logits (e.g., Softmax). The threshold θ is used to distinguish IND ($>= \theta$) and OOD ($< \theta$) according to the scores of the decision function. The typical selection of threshold value needs to ensure high accuracy (e.g., 95%) of identifying IND.

3.2. Lottery Tickets Less Overconfident

Frankle and Carbin (2019) put forward the Lottery Ticket Hypothesis (LTH). In short, a winning ticket S^+ related to the target task can be identified from a randomly initialized overparameterized network Ω, and the remaining of the network is denoted as S^-. The relationship between the posterior modeled by the overparameterized network and the posterior modeled by the winning ticket can be formulated as:

$$\begin{aligned} p(Y|X,\Omega,\xi) &= p(Y|X,S^+,S^-,\xi) \\ &= p(Y|X,S^+,\xi)\cdot\frac{p(S^-|X,Y,S^+,\xi)}{p(S^-|X,S^+,\xi)}, \end{aligned} \quad (2)$$

where ξ is metadata, including target task, training methods, datasets, etc.

Take a closer look at the right hand of Eq.(2). Given that S^+, S^- are structurally linked and jointly optimize the objective loss supervised by Y, it is crucial to note that S^- and Y are often actually not independent, but rather often establish a certain spurious positive correlation further exacerbated by the intrinsic bias brought by annotation in the training set, which can be expressed as:

$$p(S^-|X,Y,S^+,\xi) >= p(S^-|X,S^+,\xi). \quad (3)$$

We also provide a heuristic proof below.

According to Bayes' theorem, the $p(S^-|X,S^+,\xi)$ can be calculated as follows:

$$p(S^-|X,S^+,\xi) = \sum_{T\in\mathcal{T}} p(S^-|X,T,S^+,\xi)p(T|X,S^+,\xi), \quad (4)$$

where $\mathcal{T}$ is the space of target tasks related to dataset D contained in ξ, X is the input space of samples in D, Y is the specific target task defined by D ($Y \in \mathcal{T}$). The definition of parameters remains consistent with the previous context.

Since D is generally collected for a specific type target task, i.g.,Y, X could not act on other type tasks $T \in \mathcal{T}-\{Y\}$ and S^+ is a subnetwork defined by Y according to the previous condtions, it can be inferred as follows:

$$\forall T \in \mathcal{T}-\{Y\}, p(T|X,S^+,\xi) \to 0. \quad (5)$$

Therefore, take Eq.(5) into Eq.(4) to get the following expression:

$$\begin{aligned} p(S^-|*) &= \underbrace{p(S^-|X,Y,S^+,\xi)p(Y|X,S^+,\xi)}_{Y\in\mathcal{T}} \\ &+ \underbrace{0+\cdots+0+\cdots}_{Y\in\mathcal{T}-\{Y\}} \\ &= p(S^-|X,Y,S^+,\xi)p(Y|X,S^+,\xi) \end{aligned} \quad (6)$$

According to the Eq.(6), the following can be obtained:

$$\begin{aligned} p(S^-|X,Y,S^+,\xi) &= \frac{p(S^-|X,S^+,\xi)}{p(Y|X,S^+,\xi)} \\ &>= p(S^-|X,S^+,\xi) \end{aligned} \quad (7)$$

Therefore, Eq.(2) can be further calculated as:

$$p(Y|X,S^+,S^-,\xi) >= p(Y|X,S^+,\xi). \quad (8)$$

The above expression shows that the overparameterized network prefers to be more overconfident than the winning ticket, which results in giving high confidence to OOD samples, making it difficult to distinguish between IND and OOD.

3.3. The Road to Open-would Lottery Tickets

It is worth noting that the origin Lottery Ticket Hypothesis is only suitable for the closed-world (i.e., the training and test set come from the same distribution). Further, we extend this hypothesis to the open-world setting–Through one-shot pruning and minor post-processing, we can find luckier winning tickets, which can not only ensure the accuracy of IND intent identification but also better detect OOD intent with the original initialization.

Backbone and IND Identification We choose Pre-Trained Model BERT (Devlin et al., 2019), represented by $\mathcal{F}(x;\theta)$ with initialization θ_0, as the backbone network. To enable the model to effectively identify IND intent, we finetune $\mathcal{F}(x;\theta)$ under the supervision of softmax cross-entropy as suggested in Zhou et al. (2022). The objective $\mathcal{L}_{ce}$ can be formed as:

$$\mathcal{L}_{\text{ce}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathcal{F}_{y_i}(z_i))}{\sum_{j=1}^{[Y]}\exp(\mathcal{F}_j(z_i))}, \quad (9)$$

where y_i is the label of sample z_i, $\mathcal{F}_j(z_i)$ denotes the logit of the j^{th} class and θ denotes parameters.

Seek Parameters Need to be Masked Different from Iterative Magnitude Pruning (IMP), we generate lottery tickets through one-shot pruning. To this end, inspired by Louizos et al. (2017), we also add a binary "gate" to each parameter in the model to determine whether the parameter is of interest to the target task. Specifically, with a Pre-Trained Model $\mathcal{F}(x;\theta)$ at hand, the subnetwork is generated by $\mathcal{F}(x;\theta\odot M)$, where $M\in\{0,1\}^{|\theta|}$ denotes the "gates"and $|\theta|$ is the size of the parameters.

However, due to M being a discrete (non-differentiable) binary and the exponential combinatorial property of $2^{|\theta|}$, it cannot be optimized normally. Following Louizos et al. (2017), we put Bernoulli distribution over the entry $m_i\sim Bern(\pi_i)$, where $m_i\in M$ and $\pi_i = Pr(m_i=1)$. In addition to the convenience of optimization, the purpose of introducing random variables is that if the probability of a parameter, i.e., π_i is too small, we can consider that the parameter is not strongly related to the task, which means it can be "masked". The framework of optimization can be defined as:

$$\mathcal{L}_{\text{mask}}(\boldsymbol{\pi}) = \mathsf{E}_{q(M;\pi)}[\mathcal{L}_{ce}(\theta\odot M)] + \mathcal{R}(\boldsymbol{\pi}) \quad (10)$$

where $\mathcal{L}_{ce}$ stands for above loss in Eq. (9) but the parameter to be optimized has changed from θ to π. The $\mathcal{R}(\boldsymbol{\pi})$ is a regularization term w.r.t. parameters $\boldsymbol{\pi}$. The regularization term can have different forms for different purposes. Here we adopt L_0 regularization to encourage sparsity.

To further optimize the first term of Eq. (10) (which cannot be optimized based on gradient due to the discrete nature of M), following Louizos et al. (2017), we "smooth" the Eq. (10). With the help of the uniform distribution $\mathcal{U}(0,1)$ and the binary concrete continuous random variable s_i which is distributed in the (0,1) interval, we can reparameterize ($\mathcal{H}$) the M and an entry $m_i\in M$ can be reparameterized as follows:

$$u_i\sim\mathcal{U}(0,1), \quad (11)$$
$$s_i = \text{Sigmoid}((\log u_i/(1-u_i)+\alpha_i)/\beta), \quad (12)$$
$$m_i = \min(1,\max(0,s_i(\zeta-\gamma)+\gamma)), \quad (13)$$

where (α_i,β) are the parameters of the binary concrete distribution and $(\zeta<0,\gamma>1)$ are constants to stretch the distribution interval of s_i. Then objective of Eq.(10) can be rewritten as:

$$\mathcal{L}_{mask} = \mathsf{E}_{u\in\mathcal{U}(0,1)}[\mathcal{L}_{ce}(\theta\odot\mathcal{H}(u,\boldsymbol{\alpha}))] + \lambda\cdot\sum_{i=1}^{|\theta|}\text{Sigmoid}(\alpha_i-\beta\log\frac{-\zeta}{\gamma}), \quad (14)$$

where $\mathcal{H}$ is above reparameterization and λ is a hyper-parameter to balance two terms in $\mathcal{L}_{mask}$. In practice, we can adopt Monte Carlo as Louizos et al. (2017) to the expectation (i.e. the first term) due to reparameterization.

Retrain with Origin Initialization After the optimization of Eq. (14) converges or the iteration reaches a certain number of epochs, for each parameter, the associated probability π, which can be considered as the degree of correlation between the parameter and the target task, can be output, and mask can be obtained by $M=\mathbb{I}(\boldsymbol{\pi}\geq\mu)$, where $\mathbb{I}$ is indicator function and μ is the threshold to filter parameters. Finally, assign to the unmasked parameters original initial values in θ_0 and retrain the model with new initialization $\hat{\theta}_0=\theta_0\odot M$.

3.4. OOD Detection with Lottery Tickets

To better explore the ability of lottery tickets to detect OOD, we just take maximum softmax probability as scoring function and do not select relatively complex OOD scoring functions, such as *Energy*, *ReAct*, and so on (we also demonstrate our lottery tickets can be well compatible with these downstream detection functions in Section 6.2). However, we will carry out temperature scaling on the logits before that. We will further demonstrate the effectiveness of temperature scaling in theory.

What is Temperature Scaling? The temperature scaling is just a simple extension of the softmax score. Its definition is as follows:

$$\mathcal{S}_i(x;T) = \frac{\exp(\phi_i(x)/T)}{\sum_{j=1}^{k}\exp(\phi_j(x)/T)}, \quad (15)$$

where T (usually T $>$ 1) is called the temperature. In Section 6.4, we will analyze it in detail.

Why is Temperature Scaling? In addition to calibration, we observed an interesting and common phenomenon (also mentioned in the computer vision field (Hendrycks et al., 2019; Liang et al., 2018)). For a pair of indistinguishable IND and OOD samples, excluding the maximum logit score, we find that the remaining logit scores for the IND sample are more uneven (see strict measurement in following prove) than that for the OOD sample. The intrinsic lie in that the general characteristics of the intent of the IND sample are similar to another (or more) intent and significantly different from other intents. Those similar intents will be assigned high confidence, while other (different) intents would be given relatively low confidence, especially when the number of intents is large (Hendrycks et al., 2019). This will cause the confidence of ground truth intent to be dispersed by similar intents, making the model misidentify as OOD.

Different from the previous empirical demonstration, in the following proposed theorem, we theoretically demonstrate why temperature scaling (just needs to be greater than 1) can differentiate In- and Out-of-Domain based on the above properties and bring new insights into Temperature Scaling.

Theorem 3.1. *Let $x_A \in D_{IND}$ and $x_B \in D_{OOD}$ be from IND and OOD respectively, the logits outputed by pre-trained model $\mathcal{F}$ are $\phi_A = \{a_1, ..., a_k\}$ and $\phi_B = \{b_1, ..., b_k\}$ respectively. Suppose $a_1 = \max \phi_A$ and $b_1 = \max \phi_B$ and the probabilities of both are equal after softmax, i.e., $S_1(x_A; T = 1)$=$S_1(x_B; T = 1)$. Under the condition that the distribution of $\phi_A - \{a1\}$ is more uneven than that of $\phi_B - \{b1\}$, after temperature scaling, $S_1(x_A; T > 1) \geq S_1(x_B; T > 1)$.*

Proof. According to conditions ϕ_A=$\{a_1, a_2, \cdots, a_k\}$, ϕ_B=$\{b_1, b_2, \cdots, b_k\}$, and $S_1(x_A; T = 1)$=$S_1(x_B; T = 1)$ we have:

$$\frac{\exp(a_1)}{\sum_{j=1}^{k} \exp(a_j)} = \frac{\exp(b_1)}{\sum_{j=1}^{k} \exp(b_j)} \tag{16}$$

We can get equality by Eq. (16) as:

$$\mathcal{A}(\phi_A) = \mathcal{A}(\phi_B); \tag{17}$$

$$\mathcal{A}(\phi_A) = \sum_{j=2}^{k} \exp(a_j - a_1); \tag{18}$$

$$\mathcal{A}(\phi_B) = \sum_{j=2}^{k} \exp(b_j - b_1). \tag{19}$$

Now Let us consider introducing Temperature Scaling T(>1), $\mathcal{A}(\phi_A)$ and $\mathcal{A}(\phi_B)$ become as:

$$\mathcal{A}(\phi_A, T) = \sum_{j=2}^{k} (\exp(a_j - a_1))^{\frac{1}{T}}; \tag{20}$$

$$\mathcal{A}(\phi_B, T) = \sum_{j=2}^{k} (\exp(b_j - b_1))^{\frac{1}{T}}. \tag{21}$$

According to properties of inequalities in Chen (2014), $\sum_{j=2}^{k} \exp(x_j)^{\frac{1}{T}}$ is **concave** and take maximum value when $\{x_j\}$ is even (equal with each other) denoted as $\bar{X}$.
And since the distribution of $\phi_A - \{a1\}$ is more uneven than that of $\phi_B - \{b1\}$, which can be fomulated as: $\phi_A - \{a1\} \in \sigma(\bar{X})$ and $\phi_B - \{b1\}$ is out of the range of $\sigma(\bar{X})$ (σ is local space spanned by $\bar{X}$ as the center). Combining the properties of **concave**, we can get $\mathcal{A}(\phi_A, T) \leq \mathcal{A}(\phi_B, T)$ and also have:

$$\sum_{j=2}^{k} (\exp(a_j - a_1))^{\frac{1}{T}} \leq \sum_{j=2}^{k} (\exp(b_j - b_1))^{\frac{1}{T}}; \tag{22}$$

$$\frac{\exp(a_1)^{\frac{1}{T}}}{\sum_{j=1}^{k} \exp(a_j)^{\frac{1}{T}}} \geq \frac{\exp(b_1)^{\frac{1}{T}}}{\sum_{j=1}^{k} \exp(b_j)^{\frac{1}{T}}} \tag{23}$$

Then, that is $S_1(x_A; T > 1) \geq S_1(x_B; T > 1)$. □

According to the above full proof, the lead-in of temperature can make full use of such properties, which can effectively cope with such a dilemma.
Scoring function Based on the above calibrated softmax score, the definition of score function $\mathcal{G}$ we adopted is as:

$$\mathcal{G}(x, \mathcal{F}) = \max_i \{\mathcal{S}_i(x; T)\}. \tag{24}$$

When the score $\mathcal{G}$ of an utterance is less than a specific threshold θ, it can be regarded as OOD, otherwise, it is IND. As mentioned above, the selection of the threshold needs to ensure the accuracy of the IND. Refer to Section 4.2 for specific metrics.

4. Experiments

4.1. Datasets

To exhibit the effectiveness and universality of detecting OOD in lottery tickets, we extensively experiment and analysis on three used widely and challenging real-world datasets.
CLINC-FULL (Larson et al., 2019) is dataset that has been annotated and refined manually for evaluating the ability of OOD detection. It has 150 different intents covering 10 various domains and contains 22500 IND samples, 1200 OOD samples respectively.
CLINC-SMALL (Larson et al., 2019) is a variant

version of CLINC-FULL and is to measure the ability of OOD dectection of model in the case of insufficient samples. The data also has 150 intents, but each type contains only 50 samples.

StackOverflow (Xu et al., 2015) is a public corpus from Kaggle.com. The dataset involves 20 intents, in which the training set, validation set, and test set contain 12000, 2000, and 6000 samples respectively.

BANKING (Casanueva et al., 2020) It is a dataset about bank-related businesses. Its character is that the number of samples in each category of the dataset is different. The data set includes 77 different categories and the training, and test sets contain 9003, and 3080 respectively. In addition, the validation set also contains 1000 samples.

4.2. Evaluation Metrics and Baselines

For all the above datasets, we treat all OOD samples as one rejected class as following previous works (Liang et al., 2018; Zhou et al., 2022). To evaluate the performance of our method fairly, we follow previous work (Liang et al., 2018; Sun et al., 2021) and adopt two widely used metrics:

TNR at 95% TPR (TNR95) (TNR is short for ture negative rate) is to measure the probability that OOD is correctly detected correctly when the true positive rate (TPR) is up to 95%.

Area Under the Receiver Operating Charateristic curve (AUROC) is a threshold-free metric, which reflects the probability of OOD being recognized as OOD is greater than that of IND. A greater value suggests better performance.

ACCURACY (ACC) In addition, to better evaluate the overall performance of our method, that is, in addition to detecting OOD, it should effectively identify the specific class of IND. Therefore, We also introduce ACC for all categories.

We extensively compare our method with as many competitive OOD detection algorithms (scoring functions) as possible. The entire baseline can be roughly grouped into the following categories: **MSP** (Hendrycks and Gimpel, 2017), **MaxLogit** (Hendrycks et al., 2019), **Energy** (Liu et al., 2020), **Entropy** (Zheng et al., 2020) are functions of logits. **ODIN** (Liang et al., 2018) are functions of calibrated logits and **Mahalanbis distance** (Lee et al., 2018) is a function of feature. All baselines are introduced in Section 2. For a fair comparison, the network backbone (**BERT**) and training loss function (**Cross-Entropy loss**) of all methods are consistent. All methods do not use or construct additional OOD samples during the training process.

4.3. Experimental Setting

For data preprocessing, we follow previous work (Zhou et al., 2022). For the dataset Banking and Stackoverflow (the datasets do not contain a specified OOD class), we randomly select 75% of the whole intent classes as IND, get rid of other classes (remaining 25%) in the train set (also in verification set), and unify the abandon classes as OOD in the test set. For Clinc-Full and Clinc-Small, we use the specific OOD class included in the dataset itself without additional processing. During the training, we do not utilize any prior knowledge about OOD.

For the network backbone, we use the BERT (bert-uncased, with 12-layer transformer block) provided by Huggingface Transformers. The parameters we used are also widely recommended. We used an AdamW optimizer with a batch size of 32 and tried learning late in $\{1e-5, 2e-5, 5e-5\}$. In the finetune stage, we trained BERT for 30 epochs. During retraining subnetwork, we tried epochs in $\{15, 20, 30\}$(less than epochs in finetuning). In practice, satisfactory performance can be achieved by just masking the parameters of specific layers. To train efficiently and achieve better performance, we introduce a hyper-parameter to help specify which layer parameters need to be masked. All experiments are conducted in the Nvidia GeForce RTX-2080 Graphical Card with 11G graphical memory.

5. Main Results

Main Results Table 1 shows the comparison of the lottery ticket uncovered from **BERT** and other competitive OOD detection methods on different datasets. The highlighted results are the best and demonstrate our method can be better than other methods on different datasets and metrics. The results also show that our method can not only ensure the identification of IND but also detect OOD more effectively. At the same time, it can be seen from the above baselines that some detection methods, such as *Energy*, are very competitive. In subsequent experiments, we found that the combination of the lottery ticket and these methods can also achieve better results than the original, which further verifies our proposed Open-world Lottery Ticket Hypothesis. All reported results are average by conducting at least three rounds with different seeds.

6. Analysis and Discussions

6.1. A Mask for OOD Intent Classification

In the above process of finding lottery tickets, we need to retrain the subnetwork after resetting unmasked parameters to the original initialization. Ac-

Methods	Clinc-Full				Clinc-Small			
	ACC	TNR95	AUROC	AVG.	ACC	TNR95	AUROC	AVG.
MSP	$91.48_{0.18}$	$82.27_{0.78}$	$95.68_{0.21}$	89.81	$90.19_{0.06}$	$79.10_{0.79}$	$95.01_{0.36}$	88.10
MaxLogit	$91.97_{0.12}$	$85.47_{0.53}$	$96.02_{0.26}$	91.15	$90.86_{0.01}$	$82.87_{0.33}$	$95.73_{0.46}$	89.82
Energy	$92.01_{0.18}$	$85.73_{0.74}$	$96.08_{0.26}$	91.27	$90.98_{0.13}$	$84.00_{0.22}$	$95.83_{0.47}$	90.27
Entropy	$91.61_{0.20}$	$83.07_{0.99}$	$95.96_{0.22}$	90.21	$90.70_{0.02}$	$82.13_{0.73}$	$95.41_{0.38}$	89.41
ODIN	$91.99_{0.10}$	$85.60_{0.45}$	$96.11_{0.22}$	91.23	$90.92_{0.07}$	$83.23_{0.12}$	$95.87_{0.38}$	90.01
Mahalanbis	$91.94_{0.23}$	$84.90_{1.14}$	$96.79_{0.14}$	91.21	$90.52_{0.22}$	$81.10_{0.99}$	$96.26_{0.04}$	89.29
OLT(Ours)	$\mathbf{92.30}_{0.10}$	$\mathbf{86.90}_{0.50}$	$\mathbf{96.82}_{0.14}$	**92.01**	$\mathbf{91.22}_{0.07}$	$\mathbf{84.53}_{0.25}$	$\mathbf{96.32}_{0.09}$	**90.69**

Methods	Banking				Stackoverflow			
	ACC	TNR95	AUROC	AVG.	ACC	TNR95	AUROC	AVG.
MSP	$79.25_{1.25}$	$43.73_{3.95}$	$85.42_{3.74}$	69.47	$74.95_{1.15}$	$32.62_{2.09}$	$89.66_{0.71}$	65.74
MaxLogit	$80.28_{2.28}$	$48.73_{7.96}$	$86.40_{1.41}$	71.80	$75.25_{1.27}$	$33.47_{2.38}$	$90.08_{0.94}$	66.27
Energy	$79.71_{3.00}$	$47.11_{10.45}$	$86.07_{1.73}$	70.96	$75.01_{1.41}$	$32.58_{3.15}$	$90.11_{0.97}$	65.90
Entropy	$80.18_{1.29}$	$47.67_{3.92}$	$85.95_{3.61}$	71.27	$75.36_{0.98}$	$33.85_{1.26}$	$89.93_{0.65}$	66.38
ODIN	$80.33_{2.46}$	$49.12_{8.70}$	$86.33_{1.62}$	71.93	$75.30_{1.14}$	$33.56_{1.83}$	$90.36_{0.55}$	66.41
Mahalanbis	$78.84_{1.71}$	$43.20_{4.95}$	$88.31_{2.20}$	70.12	$75.19_{0.41}$	$33.62_{0.97}$	$90.71_{0.78}$	66.51
OLT(Ours)	$\mathbf{82.89}_{0.94}$	$\mathbf{58.51}_{3.89}$	$\mathbf{89.26}_{1.52}$	**76.89**	$\mathbf{75.92}_{1.16}$	$\mathbf{35.53}_{2.28}$	$\mathbf{91.36}_{0.34}$	**67.60**

Table 1: **Main Results** of comparison between Open-world Lottery Ticket (OLT) and other competitive OOD detection algorithms. **ACC** is used to measure the overall performance of the model, including both OOD detection and the identification of IND specific class. All reported results are percentages and mean by conducting with different seeds (The subscripts are the corresponding standard deviations).

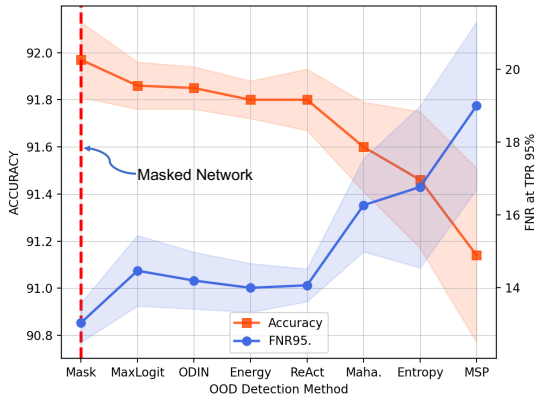


Figure 2: Masked Network vs. Baselines (Clinc-Full). The red dotted line marks the performance of masked network (subnetwork). Left Y-axis represents the accuary and Right Y-axis represents the FNR@95%TPR (The lower the value, the better).

cording to the previous analysis, the subnetwork could also be better calibrated than the original network. Can we get competitive results without retraining? We have also explored this. As suggested in Louizos et al. (2017), we obtain mask M and an entry $m_i \in M$ calculated as:

$$m_i = \min(1, \max(0, \sigma(\log \alpha_i)(\zeta - \gamma) + \gamma)), \quad (25)$$

where α_i is the parameters in random variable s_i in Eq. (12) , σ is Sigmoid function. This operation can also be regarded as $\mathcal{L}_0$ norm regularization constraint on parameters and see Louizos et al. (2017) for details.

As shown in Fig 2, we can also uncover a masked network (subnetwork) that can effectively detect OOD while maintaining the performance of IND identification. (For a clearer demonstration, we adopt FNR95 metric here to measure OOD detection ability. The lower the value, the better). [3]The experimental results are consistent with our previous claim.

Furthermore, masking without retraining can also achieve satisfactory results. **Is retraining necessary?** Our preferred answer is necessary. We surmise that retraining can learn parameters that are more suitable for the structure of the subnetwork. We hope that our experiments can inspire further theoretical or empirical research.

6.2. Towards Open-world Lottery Ticket

To further verify the versatility of the open-world lottery ticket and our extension to the LTH, we demonstrate that the gain of the effect originates from the calibrated network itself. Therefore, we combine

[3]FNR95 is to measure the probability that OOD is wrongly detected when the TPR is up to 95%.

Methods	Banking		Stackoverflow	
	ACC	Auroc	ACC	Auroc
MaxLogit	$80.28_{2.28}$	$86.40_{1.41}$	$75.25_{1.27}$	$90.08_{0.94}$
OLT+MaxLogit	$\mathbf{83.03}_{0.96}$	$\mathbf{89.54}_{1.67}$	$\mathbf{75.92}_{1.13}$	$\mathbf{91.32}_{0.39}$
Energy	$79.71_{3.00}$	$86.07_{1.73}$	$75.01_{1.41}$	$90.11_{0.97}$
OLT+Energy	$\mathbf{82.66}_{1.35}$	$\mathbf{89.27}_{1.88}$	$\mathbf{75.99}_{1.08}$	$\mathbf{91.38}_{0.34}$
Entropy	$80.18_{1.29}$	$85.95_{3.61}$	$75.36_{0.98}$	$89.93_{0.65}$
OLT+Entropy	$\mathbf{82.23}_{0.53}$	$\mathbf{87.28}_{1.44}$	$\mathbf{76.11}_{1.06}$	$\mathbf{91.08}_{0.45}$
ODIN	$80.33_{2.46}$	$86.33_{1.62}$	$75.30_{1.14}$	$90.36_{0.55}$
OLT+ODIN	$\mathbf{82.89}_{0.94}$	$\mathbf{89.26}_{1.52}$	$\mathbf{75.92}_{1.16}$	$\mathbf{91.36}_{0.34}$
Mahalabobis	$78.84_{1.71}$	$88.31_{2.20}$	$75.19_{0.41}$	$90.71_{0.78}$
OLT+Maha.	$\mathbf{81.18}_{1.30}$	$\mathbf{90.34}_{1.31}$	$\mathbf{76.00}_{0.69}$	$\mathbf{91.21}_{0.35}$

Table 2: The Lottery ticket with various OOD Scoring. **OLT** denotes the backbone is the Open-world Lottery Ticket. Shadow represents our results.

the lottery ticket with the various OOD scoring functions and compare performances with the original network. The results are shown in Table 2. From the above results, it can be seen that since the lottery ticket provides calibrated confidence, it can be more compatible with different downstream OOD detection functions and can better differentiate the distribution IND and OOD (showing the value of Auroc is high), especially those related to softmax, such as *Energy* and *MaxLogit*. At the same time, due to differentiation, the lottery network can also better maintain the identification of IND to achieve a higher overall performance (showing the value of ACC is high). These experimental results are consistent with our expectations and can be used as the basis for the establishment of the Open-world Lottery Ticket Hypothesis.

Methods	Banking		Stackoverflow	
	ACC	TNR95	ACC	TNR95
MaxLogit	$76.51_{0.48}$	$41.14_{4.16}$	$72.81_{0.53}$	$32.09_{2.27}$
OLT+MaxLogit	$\mathbf{76.65}_{0.94}$	$\mathbf{43.02}_{2.16}$	$\mathbf{73.19}_{1.07}$	$\mathbf{32.31}_{3.74}$
Energy	$\mathbf{76.38}_{0.45}$	$41.27_{4.14}$	$72.95_{0.72}$	$32.85_{2.01}$
OLT+Energy	$76.35_{1.43}$	$\mathbf{42.41}_{3.72}$	$\mathbf{73.39}_{1.26}$	$\mathbf{33.06}_{4.72}$
Entropy	$\mathbf{75.97}_{0.75}$	$\mathbf{38.90}_{5.13}$	$72.34_{0.57}$	$30.44_{2.12}$
OLT+Entropy	$75.69_{0.56}$	$38.73_{1.08}$	$\mathbf{72.96}_{0.96}$	$\mathbf{31.54}_{2.93}$
ODIN	$76.51_{0.55}$	$41.23_{4.50}$	$72.57_{0.60}$	$31.27_{2.58}$
OLT+ODIN	$\mathbf{76.71}_{0.96}$	$\mathbf{43.20}_{2.22}$	$\mathbf{73.26}_{1.08}$	$\mathbf{32.53}_{4.01}$
Mahalabobis	$7.83_{1.75}$	$5.88_{2.85}$	$72.73_{0.67}$	$31.58_{2.91}$
OLT+Maha.	$\mathbf{76.35}_{0.49}$	$\mathbf{40.70}_{2.48}$	$\mathbf{73.24}_{0.74}$	$\mathbf{32.62}_{4.22}$

Table 3: The Open-world Lottery ticket identified from RoBERTa with various OOD scoring functions.

6.3. Generality of Open-World Lottery Ticket

In Section 6.2, we have empirically verified Open-world the Lottery Ticket Hypothesis in BERT. In this section, we explore the generality of the Open-world Lottery Ticket Hypothesis and take RoBERTa (Liu et al., 2019) as an example to verify whether the Open-world Lottery Ticket Hypothesis is also valid in other models. First of all, We prune a lottery network (OLT) from RoBERTa according to our proposed method in Section 3.3. Then, we replace different *post hoc* scoring functions and make a comprehensive comparison, as we did in Section 6.2. The results are shown in Table 3. From the table, we can see that the lottery network discovered from RoBERTa can be also compatible with various scoring functions. We have preliminarily verified the generality of OLTH, and we hope that the follow-up work will bring more theoretical and experimental research.

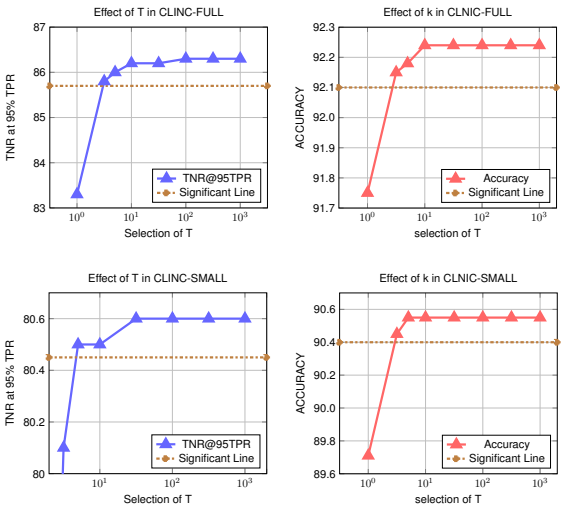


Figure 3: Effect of Temperature Scaling. As T becomes larger, the benefits brought by T will soon become smaller.

6.4. Analysis On Temperature Scaling

In Theorem 3.1, we demonstrate that temperature scaling can help differentiate the distribution between IND and OOD. Let us take a closer look at the behavior of temperature here. In the previous work (Liang et al., 2018), it is suggested to take a sufficiently larger value of temperature. However, from the proof of Theorem 3.1, it can be seen that temperature just needs to be greater than 1. We choose temperatures at different scales to test the effect (of temperature) on different data sets and the results are shown in Figure 3. We find that after $T > 1$ (without a large value), the effect of OOD detection is very significant. As T becomes

larger, the benefits brought by T will soon become smaller, which is in line with our expectations.

7. Conclusion and Future Work

Does the model know what it does not know? This paper makes an in-depth discussion of the theory and the practice. Firstly, we discuss the reasons that prevent the model from giving trustworthy confidence. Then, we uncover a subnetwork from an overparameterized model to provide calibrated confidence (helpful to differentiate in IND and OOD). In addition, We prove that temperature scaling can help distinguish IND and OOD. Combined with calibrated confidence of subnetwork and temperature scaling, we further extend the LTH to the open-world empirically and verify our conjecture by experiments.

In a larger scope, the research of this paper can be categorized more broadly into the knowledge boundary (or capability boundary) of models, which is a fundamental and critical issue in the deep learning field. With the unprecedented prevalence of artificial intelligence in recent years, research on the knowledge boundary of models, especially large-scale pre-trained language models, has drawn strong attention from scholars in academia and industry, and the research scope has become more extensive (Kadavath et al., 2022; Yin et al., 2023; Cheng et al., 2024).

First, in terms of model structure, existing research is increasingly focusing on large-scale generative architectures (Touvron et al., 2023). Can the open-world lottery ticket also be found in generative models? According to current research (Azaria and Mitchell, 2023), the answer seems to be affirmative. Then, from the perspective of task form, this study focuses on identifying the capability boundary of models. In practical scenarios, it is equally important to extend the capability boundary of models. For this purpose, a class of research (Zhou et al., 2023a) has extended the task paradigm by collecting corpora that are not within the capability boundary of the model after identification. These corpora can be further fine-grained discovered (Zhang et al., 2021b; Zhou et al., 2023b) to enhance the capability boundary of the model in practical scenarios. The proposed open-world lottery ticket mainly aims to enhance the cognitive boundary of the model. How can it further adapt to the extension of the model's capability?

Finally, as the parameter scale of models becomes increasingly larger, the inference speed of models is gradually becoming a bottleneck for their application in practical scenarios. Existing work (Zhou et al., 2023c) has preliminarily discovered that model inference optimization can not only improve the inference speed but also maintain their overall performance. How to perform inference optimization based on the *Open-World Lottery Ticket Hypothesis* is also a direction worth paying attention to.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No.2022ZD0160102).

8. Bibliographical References

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

Steven Cao, Victor Sanh, and Alexander Rush. 2021a. Low-complexity probing via finding subnetworks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–966, Online. Association for Computational Linguistics.

Steven Cao, Victor Sanh, and Alexander M. Rush. 2021b. Low-complexity probing via finding subnetworks. *CoRR*, abs/2104.03514.

Inigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Evan Chen. 2014. A brief introduction to olympiad inequalities.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Kai Chen, and Xipeng Qiu. 2024.

Can ai assistants know what they don't know? *ArXiv*, abs/2401.13275.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. *CoRR*, abs/1706.04599.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. 2019. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *CoRR*, abs/2207.05221.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1311–1316. Association for Computational Linguistics.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *CoRR*, abs/1807.03888.

Shiyu Liang, Yixuan Li, and R. Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *CoRR*, abs/2010.03759.

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. *CoRR*, abs/1907.11692.

Christos Louizos, Max Welling, and Diederik P. Kingma. 2017. Learning sparse neural networks through l_0 regularization. *CoRR*, abs/1712.01312.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 427–436. IEEE Computer Society.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

Yiyou Sun, Chuan Guo, and Yixuan Li. 2021. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 144–157.

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Zijun Liu, Yanan Wu, Hong Xu, Huixing Jiang, and Weiran Xu. 2021. Modeling discriminative representations for out-of-domain detection with supervised contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 870–878. Association for Computational Linguistics.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert Y. S. Lam. 2021. Out-of-scope intent detection with self-supervision and discriminative training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3521–3532. Association for Computational Linguistics.

Dinghuai Zhang, Kartik Ahuja, Yilun Xu, Yisen Wang, and Aaron C. Courville. 2021a. Can subnetwork structure be the key to out-of-distribution generalization? In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12356–12367. PMLR.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021b. Discovering new intents with deep aligned clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14365–14373.

Rui Zheng, Bao Rong, Yuhao Zhou, Di Liang, Sirui Wang, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Robust lottery tickets for pre-trained language models. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2211–2224, Dublin, Ireland. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020. Out-of-domain detection for natural language understanding in dialog systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1198–1209.

Yunhua Zhou, Jiawei Hong, and Xipeng Qiu. 2023a. Towards open environment intent prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2226–2240, Toronto, Canada. Association for Computational Linguistics.

Yunhua Zhou, Peiju Liu, and Xipeng Qiu. 2022. KNN-contrastive learning for out-of-domain intent classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5129–5141, Dublin, Ireland. Association for Computational Linguistics.

Yunhua Zhou, Guofeng Quan, and Xipeng Qiu. 2023b. A probabilistic framework for discovering new intents. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3771–3784, Toronto, Canada. Association for Computational Linguistics.

Yunhua Zhou, Jianqiang Yang, Pengyu Wang, and Xipeng Qiu. 2023c. Two birds one stone: Dynamic ensemble for OOD intent classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10659–10673, Toronto, Canada. Association for Computational Linguistics.