ORTicket: Let One Robust BERT Ticket Transfer across Different Tasks

Yuhao Zhou[1*], **Wenxiang Chen**[1*], **Rui Zheng**[1], **Zhiheng Xi**[1],
Tao Gui[2†], **Qi Zhang**[1], **Xuanjing Huang**[3,1†]
[1] School of Computer Science, Fudan University
[2] Institute of Modern Languages and Linguistics, Fudan University
[3] International Human Phenome Institutes (Shanghai)
{zhouyh21,chenwx23}@m.fudan.edu.cn, {tgui,xjhuang}@fudan.edu.cn

Abstract

Pretrained language models can be applied for various downstream tasks but are susceptible to subtle perturbations. Most adversarial defense methods often introduce adversarial training during the fine-tuning phase to enhance empirical robustness. However, the repeated execution of adversarial training hinders training efficiency when transitioning to different tasks. In this paper, we explore the transferability of robustness within subnetworks and leverage this insight to introduce a novel adversarial defense method **ORTicket**, eliminating the need for separate adversarial training across diverse downstream tasks. Specifically, (i) pruning the full model using the MLM task (the same task employed for BERT pretraining) yields a task-agnostic robust subnetwork(i.e., winning ticket in *Lottery Ticket Hypothesis*); and (ii) fine-tuning this subnetwork for downstream tasks. Extensive experiments demonstrate that our approach achieves comparable robustness to other defense methods while retaining the efficiency of traditional fine-tuning.This also confirms the significance of selecting MLM task for identifying the transferable robust subnetwork. Furthermore, our method is orthogonal to other adversarial training approaches, indicating the potential for further enhancement of model robustness.

Keywords: language model, adversarial defense, lottery ticket hypothesis

1. Introduction

Pre-trained language models (PLMs), such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have achieved great success in the field of natural language understanding. Following self-supervised pre-training on large amounts of data, these PLMs can achieve superior performance on a wide range of downstream tasks through fine-tuning. Such pre-training and then fine-tuning paradigm significantly enhances the model's utility in downstream tasks, particularly in terms of training time and storage efficiency.

At the same time, PLMs' vulnerability to adversarial examples has been gradually revealed. A variety of well-designed adversarial attack methods prove themselves on many downstream tasks, threatening the robustness of models prevalently (Jin et al., 2020; Li et al., 2019; Li et al., 2020). To improve the empirical robustness while retaining high accuracy on clean datasets, various adversarial defense methods often require the introduction of adversarial training during the fine-tuning phase (Madry et al., 2018; Zhu et al., 2020; Li and Qiu, 2021; Wang et al., 2021a). However, the time-consuming and memory-intensive nature of adversarial training increases the cost of using these methods in downstream tasks, especially when transitioning between different tasks.

Recently proposed *Robust Lottery Ticket Hypothesis* suggests the existence of winning tickets (i.e., subnetworks) within dense networks corresponding to downstream tasks (Fu et al., 2021; Zheng et al., 2022). These winning tickets can achieve comparable accuracy and superior robustness compared to the full network. Furthermore, some works related to *Lottery Ticket Hypothesis* have shown structural similarity or transferability of winning tickets across different downstream tasks (Chen et al., 2020; Zheng et al., 2022; Xi et al., 2023). This implies the presence of a robust super winning ticket within pre-trained models that is independent of specific tasks and can exhibit robustness across various downstream tasks.

In this paper, we identify such task-agnostic robust winning tickets and propose a novel adversarial defense method. We employ importance-based structured pruning to the self-attention heads and intermediate neurons, removing portions that contribute less to the robustness of MLM tasks and obtaining task-agnostic robust tickets for BERT models. For various downstream tasks, fine-tuning the same winning ticket (i.e., the same subnetwork checkpoint) suffices to attain robustness. Experimental results demonstrate that our adversarial defense method empowers models to exhibit state-of-the-art robustness across many downstream tasks, while preserving the training efficiency of traditional fine-tuning. At the same time, we validate the importance of MLM tasks in the search

[*] Equal contribution.
[†] Corresponding authors.

for task-agnostic robust tickets. Furthermore, if adversarial training is applied during the fine-tuning stage of robust tickets, the model's robustness can be further enhanced, indicating the orthogonality of our method with other adversarial training techniques. Our codes are publicly available at $Github$[1].

The main contributions of our work are summarized as follows:

- We validate the robust transferability of lottery ticket networks and introduce a method to identify robust winning tickets with transferability across different tasks.
- Leveraging this task-agnostic robust ticket, we propose an efficient adversarial defense method that achieves state-of-the-art robustness across various tasks while maintaining accuracy and training efficiency. Notably, this method is orthogonal to other adversarial training techniques.
- We make the data for these task-agnostic robust tickets publicly available. This provides the open-source community with convenient access to robust models for various tasks, similar to the checkpoints for pre-trained BERT models.

2. Methodology

2.1. Revisiting Robust Lottery Ticket Hypothesis

For a network $f(x;\theta_0)$ initialized with parameters θ_0, a subnetwork is a network $f(x;\mathbf{m}\odot\theta_0)$ with a binary pruning mask $\mathbf{m}\in\{0,1\}^d$ (where $\odot$ is the Hadamard product operator and d is the dimension of θ_0). According to *Robust Lottery Ticket Hypothesis* proposed by Zheng et al. (2022), a subnetwork can be defined as a robust winning ticket of the full network if it achieves the same performance and better robustness compared with the full network after undergoing the same training process. However, the robust winning tickets obtained through this approach are unstructured and task-specific.

2.2. Task-agnostic Robust Tickets

Our goal is to extract a task-agnostic robust winning ticket on pre-trained BERT models. In retrospect, BERT uses two unsupervised tasks for pre-training, one is masked language modeling (MLM) and the other is next sentence prediction (NSP)(Devlin et al., 2019). But Liu et al. (2019) confirm that removing the NSP loss matches or slightly improves downstream task performance, which means that the MLM task contributes the majority of the powerful language modeling capabilities of BERT. Based on this observation, we resort to the MLM task to obtain task-agnostic robust tickets that can be used across various downstream tasks.

```
Algorithm 1: Searching on the MLM task
Input: model parameters $\theta_0$,
       the learning rate $\eta_0$ for the MLM task
Output: learnable importance coefficients $\mathbf{c}$
$\theta \leftarrow \theta_0$, $\mathbf{c} \leftarrow 1$;
repeat
    $\theta = \theta - \eta_0 \nabla_\theta(\mathcal{L}_{adv}(\theta, \mathbf{c}) + \mathcal{R}(\mathbf{c}))$;
    $\mathbf{c} = \mathbf{c} - \eta_0 \nabla_\mathbf{c}(\mathcal{L}_{adv}(\theta, \mathbf{c}) + \mathcal{R}(\mathbf{c}))$;
until the convergence condition in Sec.2.2.2 is
  satisfied;
return $\mathbf{c}$;
```

2.2.1. The Original Architecture of BERT

BERT (Devlin et al., 2019) is a representative of pre-trained language models in the field of natural language understading, which is a multi-layer bidirectional Transformer encoder (Vaswani et al., 2017). All layers have identical structure: a multi-head self-attention (MHA) block followed by a feed-forward network (FFN), with residual connections around each. In each layer, the MHA consists of N_h independently parameterized heads. An attention head h is parametrized by $W_K^h, W_Q^h, W_V^h \in \mathbb{R}^{d_h\times d}$, $W_O^h \in \mathbb{R}^{d\times d_h}$, which represent the matrix of query, key, value and output respectively. d is the dimension of input vectors, namely, the hidden size (*e.g.*, 768) and d_h is the dimension of the output of each head (typically set to d/N_h, *e.g.*, 64). Given an input $x\in\mathbb{R}^{d_h\times d}$ followed by the output:

$$\mathrm{MHA}_{ori}(x) = \sum_{h=1}^{N_h} \mathrm{Att}_{W_K^h, W_Q^h, W_V^h, W_O^h}(x). \quad (1)$$

The FNN consists of two linear layers which receives vectors $z\in\mathbb{R}^d$ from the attention sublayer:

$$\mathrm{FFN}_{ori}(z) = \mathrm{GELU}(zW_1 + b_1)\cdot W_2 + b_2, \quad (2)$$

where $W_1\in\mathbb{R}^{d\times d_f}$, $W_2\in\mathbb{R}^{d_f\times d}$, and $d_f = 4d$.

2.2.2. Searching Stage

Learnable Importance Coefficients Recent research reveals that self-attention heads in Transformer are redundant (Michel et al., 2019; Voita et al., 2019). We adopt the pruning method proposed by Prasanna et al. (2020), which extends the importance-based structured pruning method to BERT. Thus, the new form of MHA and FFN

[1] https://github.com/CiaranZhou/ORTicket

becomes:

$$\mathrm{MHA}(x) = \sum_{h=1}^{N_h} c_{\text{H}}^{h} \cdot \mathrm{Att}_{W_K^h, W_Q^h, W_V^h, W_O^h}(x), \tag{3}$$

$$\mathrm{FFN}(z) = c_{\text{F}} \cdot \mathrm{FFN}_{ori}(z), \tag{4}$$

where c_{H}^{h} and c_{F} denote the coefficients for the self-attention head h and the FFN respectively. Also, a regularizer is needed to limit the importance coefficients:

$$\mathcal{R}(\mathbf{c}) = \lambda_{\text{H}} \|\mathbf{c}_{\text{H}}\|_1 + \lambda_{\text{F}} \|\mathbf{c}_{\text{F}}\|_1, \tag{5}$$

where $\mathbf{c} = \{\mathbf{c}_{\text{H}}, \mathbf{c}_{\text{F}}\}$, λ_{H} and λ_{F} denote regularization strength for these two coefficients respectively.

Adversarial Loss Objective To identify the task-agnostic robust tickets, we perform adversarial training on the MLM task and introduce an adversarial loss objective:

$$\min_{\theta, \mathbf{c}} \underbrace{\mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{\|\delta\| \le \epsilon} \mathcal{L}\big(f(x+\delta; \theta, \mathbf{c}), y\big)}_{\mathcal{L}_{adv}(\theta, \mathbf{c})}, \tag{6}$$

where input-output pairs (x, y) come from training dataset $\mathcal{D}$, δ is the perturbation that is constrained within the ϵ ball, and $\mathcal{L}$ is the cross entropy loss function. The inner maximization can be solved by running a number of projected gradient descent steps (Madry et al., 2018).

Then, the final loss target is obtained by combining equation (5) and (6):

$$\min_{\theta, \mathbf{c}} \mathcal{L}_{adv}(\theta, \mathbf{c}) + \mathcal{R}(\mathbf{c}). \tag{7}$$

Early-stopping Strategy You et al. (2020) and Xi et al. (2022) discover that the winning tickets can be identified in the early training phase and that the normalized Hamming distance for $\mathbf{c}$ between consecutive miniepochs (1miniepoch = 0.05epoch) is essentially monotonically decreasing. They introduce a threshold γ to detect when to end the ticket searching stage. We follow their approach to accelerate the process of adversarial training and apply it to both MHA and FFN. To ensure the continuity of convergence, we simply set it to end the searching stage when the normalized Hamming distance is less than the threshold 5 times in a row.

After training on the MLM task, we obtain the converged importance coefficients $\mathbf{c}$. Then, We use them to draw robust winning tickets for downstream tasks, and fine-tune these tickets to obtain robustness-enhanced models after re-initializing the model weights.

2.3. Adversarial Defense with ORTicket

```
Algorithm 2: Drawing and fine-tuning
Input: learnable importance coefficients $\mathbf{c}$, model
       parameters $\theta_0$, fine-tuning epoch $N$,
       learning rate $\eta$ for fine-tuning, pruning
       ratios $\tau_{\text{H}}$ and $\tau_{\text{F}}$ for MHA and FFN
Output: Robust Ticket parameters $\theta_{ticket}$
$\mathbf{m} \leftarrow \mathbb{1}$, $c_{\text{H}}^* \leftarrow$ the $\lceil \tau_{\text{H}} N_{\text{H}} \rceil$-th smallest element in
  $\mathbf{c}_{\text{H}}$, $c_{\text{F}}^* \leftarrow$ the $\lceil \tau_{\text{F}} N_{\text{F}} \rceil$-th smallest element in $\mathbf{c}_{\text{F}}$;
foreach $\mathrm{H}_1 ... \mathrm{H}_{N_{\text{H}}}, \mathrm{F}_1 ... \mathrm{F}_{N_{\text{F}}}$ do
    if $c_{\text{H}_i} \le c_{\text{H}}^*$ then $m_{\text{H}_i} = 0$ ;
    if $c_{\text{F}_j} \le c_{\text{F}}^*$ then $m_{\text{F}_j} = 0$ ;
end
$\theta \leftarrow \theta_0$, $\theta_{ticket} \leftarrow \mathbf{m} \odot \theta_0$;
for $epoch \leftarrow 1...N$ do
    $\theta_{ticket} \leftarrow \theta_{ticket} - \eta \nabla_{\theta_{ticket}} \mathcal{L}_{CrossEntropy}$;
end
return $\theta_{ticket}$;
```

Drawing Stage and Pruning Strategy The self-attention heads and intermediate neurons with the smallest importance coefficients are considered to contribute the least to robustness. Before fine-tuning the model on the downstream tasks, it applies a binary mask $m_{\text{H}_i} \in \{0, 1\}$ to i-th MHA to prune the unimportant MHA, where $1 \le i \le N_{\text{H}}$. Then, according to the pruning ratio τ_{H}, it computes a cut-off threshold c_{H}^* as the $\lceil \tau_{\text{H}} N_{\text{H}} \rceil$-th smallest element in $\mathbf{c}_{\text{H}}$. Finally, a mask $\mathbf{m}_{\text{H}}$ is created such that $m_{\text{H}_i} = 0$ if $c_{\text{H}_i} \le c_{\text{H}}^*$ and $m_{\text{H}_i} = 1$ otherwise. Similarly, the same methodology can be applied to obtain $\mathbf{m}_{\text{F}}$. Followed Xi et al. (2022), we perform global pruning of the model instead of layer-wise pruning, and keep at least one self-attention head for each layer. For different downstream tasks, they share learned importance coefficients and differ only in the strength of pruning.

Fine-tuning Stage Fine-tuning ORTicket is the final step in the adversarial defense process. After pruning the self-attention heads and intermediate neurons, we re-initialize the model. By default, we conduct standard parameter fine-tuning on ORTicket in downstream tasks. Additionally, as the extraction of ORTicket occurs before the fine-tuning phase, it is orthogonal to adversarial training methods applied during fine-tuning. This means we can also subject ORTicket to fine-tuning with adversarial training.

2.4. A Brief Review

First of all, our adversarial defense method performs adversarial training of the pre-trained BERT models on the MLM task to show which parts of the model are robust. This constitutes the process of obtaining ORTicket, as depicted in Algorithm 1.

Then different pruning ratios are set according to different downstream tasks and the corresponding robust winning tickets are drawn. Finally we perform traditional fine-tuning of the robust tickets. This represents the process of implementing adversarial defense using ORTicket, as illustrated in Algorithm 2.

3. Experimental Settings

In this section, we conduct several experiments to demonstrate the effectiveness of our defense method over multiple NLP tasks.

3.1. Backbone and Datasets

We employ widely used BERT$_{\mathrm{BASE}}$ as the backbone model which implemented by Huggingface Transformers [2] (Wolf et al., 2020) library. It has 12 Transformer blocks, 12 self-attention heads, 3,072 intermediate neurons per layer and 109M parameters in total.

In our experiments, we consider two popular text classification datasets: Internet Movie Database (**IMDb**) (Maas et al., 2011) and AG News corpus (**AG News**) (Zhang et al., 2015). The first one is binary sentiment analysis tasks that classify reviews into positive or negative sentiment, and the other one is a classification task in which articles are categorized as world, sports, business or sci/tech.

Additionally, we introduce three tasks from the GLUE benchmark (Wang et al., 2019) as a supplement: sentiment analysis (**SST-2**), natural language inference(**QNLI**), question paring(**QQP**).

3.2. Baselines

We compare our method (**ORTicket**) against standard fine-tuning and four competitive adversarial defense methods applied during the fine-tuning stage. (1) **Vanilla** (Devlin et al., 2019): Full parameter fine-tuning on downstream tasks without employing any defensive measures. (2) **PGD** (Madry et al., 2018): Projected gradient descent formulates adversarial training algorithms into solving a min-max problem that minimizes the empirical loss on adversarial examples that can lead to maximized adversarial risk. (3) **FreeLB** (Zhu et al., 2020): An enhanced gradient-based adversarial training method which is not targeted at specific attack methods. (4) **InfoBERT** (Wang et al., 2021a): A learning framework for robust fine-tuning of PLMs from an information-theoretic perspective. (5) **RobustT** (Zheng et al., 2022): *Robust Lottery Ticket Hypothesis* finds the full PLM contains subnetworks, i.e., robust tickets, that can achieve a better robustness performance.

3.3. Attack Methods

Three well-received attack methods are adopted to evaluate our method against baselines. (1) **TextFooler** (Jin et al., 2020) identifies the words in a sentence which is important to the victim model, and then replaces them with synonyms that are semantically similar and syntactically correct until the model's prediction for that sentence changes. (2) **TextBugger** (Li et al., 2019) generates misspelled words by using character-level and word-level perturbations. (3) **BERT-Attack** (Li et al., 2020) generates adversarial samples using pre-trained masked language models exemplified by BERT, which can generate fluent and semantically preserved samples. We use TextAttack [3] toolkit to implement these attack methods in adversatial attack experiments.

3.4. Evaluation Metrics

In our experiments, we assessed the experimental methods from two perspectives: robustness and training efficiency.

Robustness Evaluation The evaluation metrics for robustness are as follows: Clean accuracy (**Clean**%) denotes the accuracy on the original test dataset. Accuracy under attack (**Aua**%) represents the accuracy under adversarial attacks. Attack success rate (**Suc**%) represents the proportion of texts successfully perturbed by an attack method out of the total number of texts attempted. Number of queries (**#Query**) refers to the average number of queries made by the attacker to the victim model. For the same attack method, models with higher robustness are expected to exhibit higher clean accuracy, accuracy under attack, number of queries, and lower attack success rate in robustness evaluations.

Training Efficiency Evaluation Training time (**Speedup**) represents the relative training speed, with the training speed of FreeLB recorded as $1\times$. Trainable parameters (**Params**) represent the number of parameters that can be optimized during the training process. All experiments are performed on the GeForce RTX 2080Ti platform.

3.5. Implementation Details

For the baseline methods, we re-implement them using their open source code and report their competitive results. **Clean**% is tested on the whole test set. **Aua**%, **Suc**% and **#Query** are evaluated on the whole test dataset for SST-2, and 1,000 randomly chosen samples for other datasets. We employ FreeLB for adversarial training in the searching stage because of its efficiency compared to PGD and its better performance compared to

[2] https://github.com/huggingface/transformers

[3] https://github.com/QData/TextAttack

Dataset	Method	Clean%	TextFooler			TextBugger			BERT-Attack		
			Aua%	Suc%	#Query	Aua%	Suc%	#Query	Aua%	Suc%	#Query
IMDb	Vanilla	92.2	28.4	69.2	1550.3	41.6	54.9	1144.7	25.3	72.6	2864.1
	PGD	**93.2**	30.2	67.6	1562.8	41.6	55.4	905.8	21.8	76.6	2114.6
	FreeLB	**93.2**	35.0	62.4	1736.9	53.0	43.1	1110.9	29.0	68.9	2588.8
	InfoBERT	$\underline{\mathbf{93.3}}$	49.6	46.8	1932.3	53.8	42.3	1070.4	47.2	49.4	3088.8
	RobustT	91.8	**58.6**	**36.2**	$\underline{\mathbf{1994.7}}$	**63.6**	**30.7**	**1153.3**	**58.0**	**36.8**	**3120.2**
	ORTicket	91.4	$\underline{\mathbf{61.6}}$	$\underline{\mathbf{32.6}}$	**1972.9**	$\underline{\mathbf{68.0}}$	$\underline{\mathbf{25.6}}$	$\underline{\mathbf{1334.1}}$	$\underline{\mathbf{60.6}}$	$\underline{\mathbf{33.7}}$	$\underline{\mathbf{3999.3}}$
AG News	Vanilla	94.6	28.6	69.8	383.3	45.2	52.2	192.5	17.6	81.4	556.0
	PGD	$\underline{\mathbf{95.0}}$	**36.8**	**61.3**	414.9	$\underline{\mathbf{56.4}}$	$\underline{\mathbf{40.6}}$	201.8	21.6	77.3	616.1
	FreeLB	$\underline{\mathbf{95.0}}$	34.8	63.4	408.5	**54.2**	**42.9**	**210.3**	20.4	78.5	596.2
	InfoBERT	94.5	33.8	64.2	395.6	49.6	47.5	194.1	**23.4**	**75.2**	**618.9**
	RobustT	**94.9**	35.2	62.9	**415.6**	49.0	48.4	206.9	21.8	77.0	617.5
	ORTicket	94.4	$\underline{\mathbf{43.0}}$	$\underline{\mathbf{54.4}}$	$\underline{\mathbf{416.1}}$	52.7	44.2	$\underline{\mathbf{266.0}}$	$\underline{\mathbf{35.7}}$	$\underline{\mathbf{62.2}}$	$\underline{\mathbf{720.0}}$
SST-2	Vanilla	91.6	8.0	91.3	97.6	32.5	64.5	53.4	29.4	67.9	49.7
	PGD	$\underline{\mathbf{92.0}}$	9.2	90.0	117.4	40.6	55.9	52.7	44.2	52.0	72.0
	FreeLB	**91.8**	**21.9**	**76.1**	**130.0**	$\underline{\mathbf{43.3}}$	$\underline{\mathbf{52.8}}$	57.1	**46.5**	**49.3**	74.1
	InfoBERT	$\underline{\mathbf{92.0}}$	13.0	85.9	116.4	41.5	54.9	52.3	43.5	52.7	70.9
	RobustT	90.9	21.0	76.9	124.2	34.6	61.9	**60.9**	43.6	52.0	**79.4**
	ORTicket	90.8	$\underline{\mathbf{29.7}}$	$\underline{\mathbf{67.3}}$	$\underline{\mathbf{135.0}}$	**42.3**	**53.4**	$\underline{\mathbf{67.3}}$	$\underline{\mathbf{46.6}}$	$\underline{\mathbf{48.7}}$	$\underline{\mathbf{79.6}}$
QNLI	Vanilla	$\underline{\mathbf{91.6}}$	5.8	93.7	182.0	10.2	88.9	112.3	18.6	79.7	174.3
	PGD	90.9	16.6	81.7	234.6	15.8	82.6	147.0	19.2	78.9	260.3
	FreeLB	91.1	21.8	76.1	**249.0**	**22.0**	**75.9**	**153.0**	**26.8**	**70.6**	**284.5**
	InfoBERT	**91.5**	20.5	77.6	244.5	15.0	83.6	119.1	19.5	78.7	209.1
	RobustT	**91.5**	**22.2**	**75.7**	231.7	21.4	76.6	125.0	26.4	71.1	260.1
	ORTicket	90.9	$\underline{\mathbf{30.9}}$	$\underline{\mathbf{66.0}}$	$\underline{\mathbf{259.1}}$	$\underline{\mathbf{29.6}}$	$\underline{\mathbf{67.4}}$	$\underline{\mathbf{157.6}}$	$\underline{\mathbf{33.3}}$	$\underline{\mathbf{63.4}}$	$\underline{\mathbf{293.7}}$
QQP	Vanilla	91.2	29.2	68.0	172.4	30.6	66.4	99.3	26.2	71.3	124.4
	PGD	91.3	17.0	81.4	164.9	22.6	75.2	83.3	33.2	63.6	132.8
	FreeLB	**91.4**	32.8	64.0	180.9	31.2	65.9	102.9	**42.6**	**53.4**	$\underline{\mathbf{212.1}}$
	InfoBERT	$\underline{\mathbf{91.9}}$	34.4	62.6	174.2	35.9	60.9	90.1	37.0	59.7	134.9
	RobustT	**91.4**	$\underline{\mathbf{37.6}}$	$\underline{\mathbf{58.9}}$	**183.8**	$\underline{\mathbf{37.8}}$	$\underline{\mathbf{58.6}}$	$\underline{\mathbf{108.3}}$	$\underline{\mathbf{42.8}}$	$\underline{\mathbf{53.2}}$	**195.7**
	ORTicket	91.0	**36.3**	**60.1**	$\underline{\mathbf{185.8}}$	**36.7**	**59.7**	**107.1**	40.6	55.4	192.4

Table 1: Experimental results of adversarial robustness evaluation on $\text{BERT}_{\text{BASE}}$. The best performance is marked in **bold and underline**; the second is marked in **bold**. A robust model should exhibit high **Aua**% and **#Query**, while maintaining low **Suc**%. Our method demonstrates matching or even higher adversarial robustness compared with baselines. It is worth noting that **ORTicket** results are obtained by fine-tuning on different downstream tasks using the same checkpoint without doing task-specific training.

FreeAT (Shafahi et al., 2019) and YOPO (Zhang et al., 2019). In the searching stage, we choose an open source dataset *wikitext-2-raw-v1* (Merity et al., 2017) and train only 1 epoch on the MLM task. We set the early-stopping threshold γ to 0.1 following Xi et al. (2022). In the drawing stage, the pruning ratio for self-attention heads is set to $1/6$. The pruning ratio for intermediate neurons is set to 0.2 or 0.3 according to the specific tasks. In the fine-tuning stage, we set the training epoch to 10 for all methods, which is a trade-off between time consumption and performance.

4. Results and Discussion

In this section, we illustrate the effectiveness and efficiency of our method with experimental results.

4.1. Main Results

Robust Evaluation The robustness evaluation results of our method and other baselines are shown in Table 1. We can observe that: (1) Across 5 tasks $\times$ 3 attack methods, ORTicket achieves high adversarial robustness and only a small drop in accuracy on clean datasets; (2) As a task-agnostic adversarial defense method, our method is not inferior to task-specific adversarial defense methods on most tasks; (3) As a structured task-agnostic lottery ticket, ORticket performs on par with non-structured task-specific RobustT.

Training Efficiency Evaluation We compare the training speed of our method with traditional fine-tuning and FreeLB on datasets of 5 different tasks, as shown in Table 2. ORTicket is $6\times \sim 8\times$ faster than the most widely used adversarial defense method, FreeLB, for training. Benefiting from the reduction of redundant trainable parameters, our method achieves robustness while being faster than traditional fine-tuning. It is also memory-efficient, which means that fine-tuning of downstream tasks can be deployed on devices with less memory.

Dataset	Method	Speedup	Params
IMDb	Vanilla	$6.4\times$	$109M$
	ORTicket	$\mathbf{8.3_{-1.5}}\times$	**87M**
AG News	Vanilla	$5.2\times$	$109M$
	ORTicket	$\mathbf{6.9_{-0.5}}\times$	**87M**
SST-2	Vanilla	$6.2\times$	$109M$
	ORTicket	$\mathbf{7.7_{-1.1}}\times$	**93M**
QNLI	Vanilla	$5.5\times$	$109M$
	ORTicket	$\mathbf{7.9_{-0.8}}\times$	**87M**
QQP	Vanilla	$6.6\times$	$109M$
	ORTicket	$\mathbf{8.1_{-0.2}}\times$	**93M**

Table 2: Training speedup results of BERT$_{\mathrm{BASE}}$ on five datasets. **Speedup** means training speedup, which is reported against adversarial training method FreeLB. The subscripts represent the case where the time consumed in the searching stage is taken into account. **Params** is the number of all trainable parameters of the model. **ORTicket** achieves sizable training accelerations compared to both vanilla and robust baselines.

Dataset	Method	Clean	Aua%		
			Fooler	Bugger	BERT
IMDb	Vanilla Ticket	91.3	25.3	32.3	35.7
	w/ FreeLB	90.9	35.7	42.3	44.0
	ORTicket	**91.4**	61.6	68.0	60.6
	w/ FreeLB	**91.4**	**73.4**	**73.6**	**69.8**
AG News	Vanilla Ticket	94.2	21.2	36.4	16.8
	w/ FreeLB	94.7	33.0	52.2	29.2
	ORTicket	94.4	43.0	52.7	35.7
	w/ FreeLB	**95.1**	**45.4**	**55.8**	**36.9**
QQP	Vanilla Ticket	90.9	27.2	28.0	33.8
	w/ FreeLB	91.0	34.4	35.2	44.4
	ORTicket	91.0	36.3	36.7	40.6
	w/ FreeLB	**91.3**	**38.6**	**38.6**	**44.8**

Table 3: Orthogonal experimental results of robustness evaluation on BERT$_{\mathrm{BASE}}$. In the searching stage, Vanilla Ticket (without adversarial training) and ORTicket (with FreeLB) can be obtained by pruning the full model. In the fine-tuning stage, adversarial training methods can be applied to the robust ticket to further improve the robustness.

4.2. Ablation Study and Orthogonality Exploration

We separately investigate the impact of the adversarial loss objective during the searching stage on the performance of ORTicket and whether the adversarial training can further enhance model robustness during the fine-tuning phase.

As shown in Table 3, the experimental results indicate the necessity of the adversarial loss objective for ORTicket, as it assists in the extraction of robust tickets during the ticket searching stage. The accuracy on clean datasets is not influenced by the introduction of the adversarial loss objective, and only decreases a little on some datasets.

As an adversarial defense method applied prior to the fine-tuning stage, ORTicket can introduce adversarial training during fine-tuning to further enhance model robustness. This demonstrates the orthogonality of ORTicket with other adversarial training methods, and how they complement each other's effects.

4.3. Different Tasks to Search for Universal Robust Tickets

In Sec.2.2, we analyze the pre-training task of BERT and empirically select the same, i.e. Masked Language Modeling task, as the task used in the transferable robust tickets searching stage. Here we investigate whether other tasks have comparable generalization capabilities to MLM tasks for downstream performance.

From Table 4, we can notice that: (1) For the same downstream task, using the MLM task in the search phase of ORTicket is even better than the task itself in most cases; (2) For the same task used to search for ORTicket, the MLM task shows better robust generalization on the downstream tasks than the other tasks; (3) The choice of tasks in the search phase of the generic robust tickets has little impact on the accuracy of the model on clean datasets on downstream tasks.

Therefore, we can draw a simple conclusion that utilizing the MLM task allows the model to exhibit not only accuracy generalization on clean datasets but also robust generalization in the face of adversarial perturbations.

4.4. Different Encoder Models

To verify whether ORTicket exists in other encoder models, we conducted a preliminary exploration using RoBERTa$_{\mathrm{BASE}}$(Liu et al., 2019). [4]

As shown in Table 5, undergoing regular fine-tuning in downstream tasks, ORTicket within RoBERTa consistently achieves superior robustness performance compared to standard fine-tuning. It even outperforms adversarial training methods in most cases.

For BERT and RoBERTa, we conducted separate analyses of their ORTickets. As shown in Figure 1, in BERT models, some intermediate layer heads are consistently pruned, while this phenomenon is alleviated in RoBERTa. This may offer an explanation for RoBERTa's superior robustness compared to BERT when considering that BERT contains

[4] Due to constraints on computational resources and time, we do not conduct an extensive hyperparameter search on RoBERTa$_{\mathrm{BASE}}$. There is still potential for improvement in our method.

	Clean%	Aua%	Clean%	Aua%	Clean%	Aua%	Clean%	Aua%	Clean%	Aua%
MLM	90.8	$\underline{\textbf{29.7}}$	90.9	$\underline{\textbf{30.9}}$	91.0	$\underline{\textbf{36.3}}$	91.4	$\underline{\textbf{61.6}}$	**94.4**	$\underline{\textbf{43.0}}$
SST-2	$\underline{\textbf{92.8}}$	**28.8**	90.8	11.6	$\underline{\textbf{91.2}}$	26.2	**91.9**	22.4	**94.4**	16.6
QNLI	90.7	10.4	**91.4**	**22.5**	90.9	**32.8**	91.2	34.8	94.2	22.0
QQP	91.3	15.0	90.8	15.8	90.9	32.6	91.8	28.0	94.3	17.0
IMDb	91.3	15.0	91.2	4.6	**91.1**	30.4	91.6	**58.2**	**94.4**	33.0
AG News	91.2	16.6	90.9	11.6	$\underline{\textbf{91.2}}$	27.2	91.8	1.4	94.2	**36.0**
BERT$_{\text{BASE}}$	**91.6**	8.0	$\underline{\textbf{91.6}}$	5.8	$\underline{\textbf{91.2}}$	29.2	$\underline{\textbf{92.2}}$	28.4	$\underline{\textbf{94.6}}$	28.6
	SST-2		**QNLI**		**QQP**		**IMDb**		**AG News**	

Table 4: Different tasks to search for transferable robust tickets. The best performance is marked in $\underline{\textbf{bold and underline}}$; the second is marked in **bold**. **BERT$_{\text{BASE}}$** represents the results of traditional fine-tuning using BERT$_{\text{BASE}}$ on different downstream tasks. The left column is the searching source task and the bottom column is the destination downstream task. The robust ticket extracted by MLM task (i.e., ORTicket) shows better generalization on various downstream tasks than the other tasks.

Dataset	Method	Clean%	Aua%		
			Fooler	Bugger	BERT
IMDb	Vanilla	93.7	39.9	50.4	33.2
	FreeLB	**93.9**	45.0	56.2	56.8
	ORTicket	92.5	**70.0**	**72.0**	**68.2**
AG News	Vanilla	94.6	39.6	53.4	31.1
	FreeLB	**95.3**	**47.6**	**61.2**	**40.2**
	ORTicket	94.8	39.8	54.8	34.0
SST-2	Vanilla	94.3	18.2	42.2	11.2
	FreeLB	**94.8**	18.6	**45.8**	13.4
	ORTicket	92.2	**23.8**	45.0	**18.4**
QNLI	Vanilla	92.0	7.3	13.3	2.3
	FreeLB	**92.6**	14.2	**21.8**	7.4
	ORTicket	91.8	**19.0**	19.2	**11.4**
QQP	Vanilla	91.1	22.6	28.6	15.7
	FreeLB	**91.8**	27.0	31.6	21.0
	ORTicket	91.4	**29.4**	**31.8**	**26.6**

Table 5: Experimental results of adversarial robustness evaluation on RoBERTa$_{\text{BASE}}$. The best performance is marked in **bold**. **ORTicket** consistently outperforms the vanilla baseline and stands on par with adversarial training method FreeLB.

more low-contributing structural components to robustness. In terms of intermediate neurons, both models exhibit that neurons in the shallow layers are more prone to pruning.

4.5. Importance of Robust Tickets Initialization and Structure

Lottery Ticket Hypothesis states that the winning ticket performs poorly out of the original initialization and the structure of one winning ticket is critical. In order to better understand which has a greater impact on the robust tickets, initialization or structure, we conduct the corresponding analytical experiments. Following Zheng et al. (2022), we avoid the effect of initialization by reinitializing the weights of robust tickets. To avoid the effect of structure and retain the effect of initialization, we use the full BERT and reinitialize the weights that are not included in the robust ticket.

Dataset	Method	Clean%	Aua%
SST-2	**ORTicket**	**90.8**	**29.7**
	w/o Initialization	80.3	1.5
	w/o Structure	90.4	8.3
QNLI	**ORTicket**	**90.6**	**30.9**
	w/o Initialization	58.8	1.9
	w/o Structure	88.8	17.6
QQP	**ORTicket**	**91.0**	**36.6**
	w/o Initialization	84.1	3.7
	w/o Structure	90.7	35.4

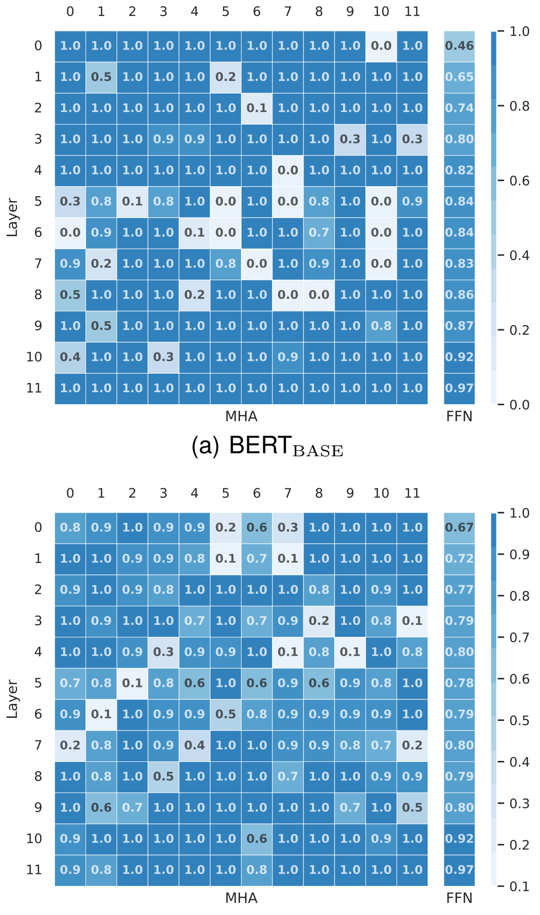
Table 6: Importance of ORTicket's initialization and structure. **Aua%** is obtained using TextFooler attack on BERT$_{\text{BASE}}$. The pre-trained initialization and the structure are both indispensable for ORTicket. The pre-trained initialization seems more important than the structure for both accuracy and robustness improvement.

Table 6 shows the importance of initialization and structure for robust tickets, as can be seen: (1) Models without the original initialization show significant degradation on both accuracy on clean datasets and robustness; (2) The structure of the robust tickets is important for robustness but not important for accuracy on clean datasets.

4.6. Impact of Pruning Ratio

In the drawing stage, a higher pruning ratio saves more memory and speeds up training, but comes with a performance degradation. We construct an experiment to investigate the trend and degree of impact of performance with pruning ratio.

As shown in Figure 2, three datasets of different task types exhibt the same trend: accuracy on the clean datasets decreases as the self-attention heads and internal neurons decrease, while the



(a) BERT$_{\mathrm{BASE}}$

(b) RoBERTa$_{\mathrm{BASE}}$

Figure 1: Heatmaps of sparsity patterns observed in BERT$_{\mathrm{BASE}}$ and RoBERTa$_{\mathrm{BASE}}$ models. In each subfigure, the cells represent the average percentage of surviving weights in self-attention heads and intermediate neurons across 10 random seeds.

robustness of the model increases and then decreases. Therefore, to extract a robust ticket with high performance, the pruning ratio should be set to about $1/6$ and $20\% \sim 30\%$ for the self-attention heads and internal neurons, respectively.

5. Related Work

Textual Adversarial Attack Textual adversarial attack methods generate adversarial examples , which are maliciously crafted by imposing small perturbations on the original input to deceive the victim model (Zhang et al., 2020). In general, textual adversarial attack methods generate character- (Eger et al., 2019), word- (Jin et al., 2020; Ren et al., 2019; Li et al., 2020), word/char- (Li et al., 2019), or sentence-level (Zhao et al., 2018) perturbations with high similarities preserved in the semantic or embedding space. These approaches can also serve as benchmarks for textual adversarial

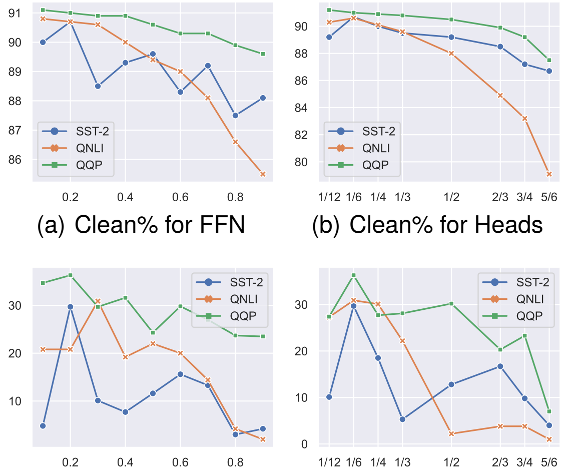


Figure 2: Impact of pruning ratio on BERT$_{\mathrm{BASE}}$ **Clean**$\%$ and **Aua**$\%$. **Aua**$\%$ is obtained under TextFooler attack. FFN and Heads refer to intermediate neurons and self-attention heads respectively. The optimal pruning ratio for ORTicket is approximately 20$\%$ for FFN and 1/6 for Heads.

defense methods.

Textual Adversarial Defense To improve the empirical robustness of models against textual adversarial attacks, more and more adversarial defense methods are proposed (Li et al., 2021; Chen et al., 2022; Liu et al., 2022). As the most popular one, adversarial training solves a min-max optimization problem (Goodfellow et al., 2015, FGSM; Miyato et al., 2017, FGM; Madry et al., 2018, PGD), which means finding the optimal adversarial perturbation (max) and optimizing the worst-case performance of the model (min).

Efficient Adversarial Training From FGSM to PGD, the quality of the generated adversarial perturbations keeps improving, but the computational cost is also increasing plus compared to the traditional training. In order to obtain the optimal adversarial perturbation faster, FreeAT (Shafahi et al., 2019) and YOPO (Zhang et al., 2019) optimize the number of gradient calculations. Later, Zhu et al. (2020) confirm that there is room for improvement in the quality of the adversarial perturbations generated by these two, and propose FreeLB, which is currently the most widely used efficient adversarial learning method in the NLP field. However, their training efficiency still falls short of conventional fine-tuning.

Lottery Ticket Hypothesis Lottery Ticket Hypothesis proposed by Frankle and Carbin (2019) suggests that the existence of matching winning tickets (i.e., subnetworks) at the initialization of the model can achieve comparable performance to the full model on downstream tasks. In the field of NLP, winning tickets are found in Transformer, LSTM, and pre-trained models such as BERT (Yu et al., 2020; Chen et al., 2020). For getting the winning tickets, different methods can be divided into structured pruning and unstructured pruning according to the pruning method. Prasanna et al. (2020) perform importance-based pruning on self-attention heads and MLPs of BERT in a structured fashion, which is efficient in the training and inference phases. Chen et al. (2020) perform unstructured magnitude pruning on all parts of the network and explores the transferability of unstructured pruned subnetworks on different downstream tasks. Fu et al. (2021) and Zheng et al. (2022) further discover the existence of robust subnetworks and propose Robust Lottery Ticket Hypothesis. To speed up the extraction process of winning tickets, Early-Bird (You et al., 2020), EarlyBERT (Chen et al., 2021) and EarlyRobust (Xi et al., 2022) extract Early-Bird tickets in the early stage of training. These studies motivate us to explore the existence of ORTicket.

6. Conclusion

In this paper, we explore the transferability of robustness within subnetworks and leverage this insight to introduce a novel adversarial defense method ORTicket, eliminating the need for separate adversarial training across diverse downstream tasks. We use a structured pruning method on the MLM task to extract a task-agnostic robust ticket for BERT models, which can obtain high robustness by fine-tuning on various downstream task. ORTicket exhibits excellent training efficiency and low training memory consumption, while remaining orthogonal to other adversarial training methods. Experiments show that our task-agnostic approach achieves robust generalization and comparable robustness to the task-specific methods on a range of downstream tasks.

7. Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No. 62206057, 61976056, 62076069, 61906176), Shanghai Rising-Star Program (23QA1400200), Natural Science Foundation of Shanghai (23ZR1403500), Program of Shanghai Academic Research Leader under grant 22XD1401100, CCF-Baidu Open Fund, and CCF-Baichuan Fund. The computations in this research were performed using the CFFF platform of Fudan University.

8. Bibliographical References

Sizhe Chen, Zhehao Huang, Qinghua Tao, Yingwen Wu, Cihang Xie, and Xiaolin Huang. 2022. Adversarial attack on attackers: Post-process to mitigate black-box score-based query attacks. *CoRR*, abs/2205.12134.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The lottery ticket hypothesis for pre-trained BERT networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xiaohan Chen, Yu Cheng, Shuohang Wang, Zhe Gan, Zhangyang Wang, and Jingjing Liu. 2021. Earlybert: Efficient BERT training via early-bird lottery tickets. 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 2195–2207. Association for Computational Linguistics.

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.

Steffen Eger, Gözde Gül Sahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding NLP systems. 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 1634–1647. Association for Computational Linguistics.

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.

Yonggan Fu, Qixuan Yu, Yang Zhang, Shang Wu, Xu Ouyang, David D. Cox, and Yingyan Lin. 2021. Drawing robust scratch tickets: Subnetworks with inborn robustness are found within randomly initialized networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 13059–13072.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025. AAAI Press.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6193–6202. Association for Computational Linguistics.

Linyang Li and Xipeng Qiu. 2021. Token-aware virtual adversarial training in natural language understanding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 8410–8418. AAAI Press.

Zongyi Li, Jianhan Xu, Jiehang Zeng, Linyang Li, Xiaoqing Zheng, Qi Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021. Searching for an effective defender: Benchmarking defense against adversarial word substitution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3137–3147. Association for Computational Linguistics.

Qin Liu, Rui Zheng, Bao Rong, Jingyi Liu, Zhihua Liu, Zhanzhan Cheng, Liang Qiao, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Flooding-x: Improving bert's resistance to adversarial attacks via loss-restricted fine-tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5634–5644. Association for Computational Linguistics.

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.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 142–150. The Association for Computer Linguistics.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Zhao Meng, Yihan Dong, Mrinmaya Sachan, and Roger Wattenhofer. 2022. Self-supervised contrastive learning with adversarial perturbations for defending word substitution-based attacks. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 87–101. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on*

Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14014–14024.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Sai Prasanna, Anna Rogers, and Anna Rumshisky. 2020. When BERT plays the lottery, all tickets are winning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3208–3229. Association for Computational Linguistics.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*, pages 1085–1097. Association for Computational Linguistics.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3353–3364.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5797–5808. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. 2021a. Infobert: Improving robustness of language models from an information theoretic perspective. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Shan Qin, Bolin Zhu, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021b. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL 2021 - System Demonstrations, Online, August 1-6, 2021*, pages 347–355. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien

Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Zhiheng Xi, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Efficient adversarial training with robust early-bird tickets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 8318–8331, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhiheng Xi, Rui Zheng, Yuansen Zhang, Xuanjing Huang, Zhongyu Wei, Minlong Peng, Mingming Sun, Qi Zhang, and Tao Gui. 2023. Connectivity patterns are task embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11993–12013. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Haoran You, Chaojian Li, Pengfei Xu, Yonggan Fu, Yue Wang, Xiaohan Chen, Richard G. Baraniuk, Zhangyang Wang, and Yingyan Lin. 2020. Drawing early-bird tickets: Toward more efficient training of deep networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Haonan Yu, Sergey Edunov, Yuandong Tian, and Ari S. Morcos. 2020. Playing the lottery with rewards and multiple languages: lottery tickets in RL and NLP. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. 2019. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 227–238.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(3):24:1–24:41.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Rui Zheng, Shihan Dou, Yuhao Zhou, Qin Liu, Tao Gui, Qi Zhang, Zhongyu Wei, Xuanjing Huang, and Menghan Zhang. 2023. Detecting adversarial samples through sharpness of loss landscape. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11282–11298. Association for Computational Linguistics.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2211–2224. Association for Computational Linguistics.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.