The Lottery Ticket Hypothesis for Vision Transformers

Xuan Shen 1 Zhenglun Kong 1 Minghai Qin 2 Peiyan Dong 1 Geng Yuan 1 Xin Meng 3 Hao Tang 4 Xiaolong Ma 5 Yanzhi Wang 1

1 Northeastern University, 2 Western Digital Research, 3 Peking university, 4 ETH Zurich, 5 Clemson University
1 shen.xu@northeastern.edu, 1 yanz.wang@northeastern.edu

Abstract
The conventional lottery ticket hypothesis (LTH) claims that there exists a sparse subnetwork within a dense neural network and a proper random initialization method, called the winning ticket, such that it can be trained from scratch to almost as good as the dense counterpart. Meanwhile, the research of LTH in vision transformers (ViTs) is scarcely evaluated. In this paper, we first show that the conventional winning ticket is hard to find at weight level of ViTs by existing methods. Then, we generalize the LTH for ViTs to input images consisting of image patches inspired by the input dependence of ViTs. That is, there exists a subset of input image patches such that a ViT can be trained from scratch by using only this subset of patches and achieve similar accuracy to the ViTs trained by using all image patches. We call this subset of input patches the winning tickets, which represent a significant amount of information in the input. Furthermore, we present a simple yet effective method to find the winning tickets in input patches for various types of ViT, including DeiT, LV-ViT, and Swin Transformers. More specifically, we use a ticket selector to generate the winning tickets based on the informativeness of patches. Meanwhile, we build another randomly selected subset of patches for comparison, and the experiments show that there is a clear difference between the performance of models trained with winning tickets and randomly selected subsets.

Introduction
In recent years, computer vision research has evolved into Transformer architectures since Vision Transformer (ViT) (Dosovitskiy et al. 2021) achieved excellent performance on image classification tasks. Models such as DeiT (Touvron et al. 2021) and Swin Transformer (Liu et al. 2021b) have achieved state-of-the-art performance on image classification on CIFAR-10/100 (Krizhevsky 2009) and ImageNet (Deng et al. 2009) datasets. As great attention has been drawn to the ViT, computation redundancy becomes an issue because of the sacrifice on lightweight model capacity for higher accuracy. Based on (Dosovitskiy et al. 2021), training a ViT from scratch requires large datasets (i.e., ImageNet), and more iterations are required for the convergence of the ViT to ensure that the network is fully exploited. The large dataset and complex model architecture motivate us to explore the lottery ticket hypothesis for efficient training and test (inference) of ViT models.

In parallel, the Lottery Ticket Hypothesis (LTH) (Frankle and Carbin 2018) suggests that there exists some winning tickets (i.e., a properly pruned subnetwork together with the original weight initialization) that can achieve performance competitive with that of the original dense network. Such winning tickets in CNNs can achieve comparable performance to the dense network on image classification tasks and allow the training of the network to be more efficient. Since then, LTH has been applied to various research areas like BERT and object detection (Chen et al. 2020; Sai Prasanna 2020; Bai et al. 2022; Shrey Desai 2019; Yu et al. 2019). Nonetheless, the existence of winning tickets of ViTs is not explored by the conventional LTH definition.

Recent works on LTH have focused on identifying the winning tickets at the weight level in DNNs. We follow the settings of conventional LTH (Frankle and Carbin 2018) in the experiments on ViTs. Interestingly, the experimental results at the weight level on DeiT-Small (Touvron et al. 2021) indicate that it is hard to find such winning tickets that make training effective even with the help of rewinding technique. Details of these experiments can be found in Table 1 and Table 2. The absence of the winning tickets is mainly due to the input-dependence of the ViT, as it is difficult to identify a subnetwork at the weight level to generalize to the different input images.

In ViTs, each image is first split into square patches, which are then projected to tokens before being fed into the transformer blocks. Recent works (Rao et al. 2021; Xu et al. 2022; Liang et al. 2022; Kong et al. 2021) demonstrate a series of dynamic token pruning strategy, where the tokens are pruned independently in the subnetworks based on the features of each input image. These observations motivate us to explore the LTH at the input level. The winning tickets are identified on the input image patches instead of model weights or tokens, and the strategy of identifying winning tickets depends on the attentiveness and informativeness of each image patch.

In this paper, we propose a new Lottery Ticket Hypothesis for Vision Transformers (LTH-ViTs), shown in Figure 1. In the first path, (a), for each input image, we use a ticket selector to produce the subset of attentive image patches before feeding them into the ViT for both training and test.
The ticket selector removes input patches with less contribution according to the attention map in transformer blocks. The positional embeddings of the remaining tokens further guide the identification of attentive image patches for training. Then, the model is trained and tested with the subset of selected patches, identified as winning tickets. The second path, (b), selects random input image patches for training. In order to compare with Path (a), we use the same ticket selector to generate input patches for test. The last path, (c), is a normal ViT training and test process where all input patches are used for training and test. The experiments we conducted in this paper show that our method of training the ViTs only with winning tickets (Path (a)) can match the performance of conventionally trained ViT models (Path (c)), and have notable accuracy advantages over ViTs trained by a random subset of input patches (Path (b)). Based on this observation, we propose the LTH for ViTs, and it suggests that there exist winning tickets at the input image patch level that can make the training and test more efficient and effective. To compensate our proposed LTH for ViTs, we also try to identify the winning tickets on CNNs with our proposed method. Interestingly, it is difficult for an input-independent CNNs to identify the winning tickets with proposed method. The results can be found in Table 4.

We summarize our contributions as follows:

- We show by experiments that the conventional winning ticket is hard to find even with the rewinding technique. In particular, there is a negligible difference in accuracy between a random sparse weight initialization and the proper initialization defined in the conventional LTH.
- We generalize the definition of winning tickets in LTH from sparse weights to sparse input. That is, there exists a subset of input image patches such that a ViT can be trained from scratch by using only this subset of patches and achieve similar accuracy to the ViTs trained by using all image patches. This subset of input patches in training is called the input winning ticket of ViTs.
- We propose a method to identify the input winning tickets for ViTs. Experiments show that we can reduce 48.8% of input image patches during training, which can match the accuracy to the ViTs trained by using all image patches. We also observe another feature of input winning tickets, which is the universality across different types of ViTs.

### Related Works

#### The Lottery Ticket Hypothesis

The lottery ticket hypothesis is first proposed by the work (Frankle and Carbin 2018). The hypothesis suggests that winning tickets found in CNN on CIFAR-10/100 (Krizhevsky 2009) dataset have connections with the initial weights that can make the training particularly effective. The following work (Renda, Frankle, and Carbin 2020) extends (rewinds) the training of the subnetwork from initialization to the early stage of pretraining, which improves the accuracy of the subnetwork in more challenging tasks. Furthermore, the work (Frankle et al. 2020) demonstrates the stability of those subnetworks to the SGD noise in the early stage of training, which explains the success of the rewinding technique in LTH. Besides, the current work (Ma et al. 2021) claims that the winning tickets are difficult to find on ImageNet (Deng et al. 2009) dataset for CNNs, and this work also shows the effect of the enough-trained original network on the identically initialized subnetwork.

Apart from the image classification tasks, the LTH is also imported in many other research areas (Chen et al. 2020; Mallya, Davis, and Lazebnik 2018; Gale, Elsen, and Hooker 2019; Yu et al. 2019; Renda, Frankle, and Carbin 2020; Chen et al. 2020; Prasanna, Rogers, and Rumshisky 2020; Girish et al. 2021). The works (Gale, Elsen, and Hooker 2019; Yu et al. 2019; Renda, Frankle, and Carbin 2020) show that the subnetworks exist early in the training instead of initialization on Transformers. Also, the works (Chen et al. 2020; Prasanna, Rogers, and Rumshisky 2020) study the LTH in BERT models and find the matching subnetworks at high-level sparsity on the pretrained initialization. Besides, the work (Girish et al. 2021) proposes a guidance for the identification of task-specific winning tickets for object detection, instance segmentation, and keypoint estimation. However, none of the existing LTH literatures explored sparse winning tickets for ViTs which have been widely utilized in computer vision tasks in recent years.

#### Redundancy in ViTs

The work (Dosovitskiy et al. 2021) first investigates the Transformers in computer vision and then builds the Vision Transformer (ViT) by dividing one image into several patches. Since then, many variants have been proposed (Wu et al. 2021; Lu et al. 2021; Song et al. 2021; Li et al. 2021; Zhou et al. 2021).
posed winning tickets are different from the conventional ones. The winning tickets in our proposed special lottery ticket hypothesis are proved to be effective according to the experiments. Furthermore, the remaining attentive image patches from training can still match the accuracy of the pretrain model. Therefore, based on these works, the token redundancy in ViTs proves to be large and each token can be connected to the corresponding input image patch. However, these approaches typically need to progressively reduce the number of tokens, because ViTs are considered difficult to implement. DynamicViT (Rao et al. 2021) introduces a method to remove the token at the very beginning due to the loss of performance. In our work, we propose an effective framework to remove the inattentive image patches before feeding them into the transformer blocks. Our work is different from the token pruning works that feed the transformer with a full image and prune the tokens between the transformer blocks progressively. The results show that removing inattentive image patches from training can still match the accuracy of the pretrained model. Furthermore, the remaining attentive image patches are proved to be effective according to the experimental results of randomness counterpart, and defined as winning tickets in our proposed special lottery ticket hypothesis in ViTs. This is the first time that the LTH achieves positive results in vision transformer architectures. The proposed winning tickets are different from the conventional LTH whose winning tickets are subnetworks.

The Lottery Ticket Hypothesis for Vision Transformers

Our work is mainly based on the DeiT (Touvron et al. 2021). We first show that it is hard to find conventional winning ticket in sparse ViTs even with rewinding technique, and then present how to identify the winning tickets with the analysis of the Transformer architectures. The rigorous definition of the lottery ticket hypothesis is elaborated and the specific methods are illustrated to demonstrate the existence and effectiveness of the winning tickets in ViTs. Furthermore, other variants of ViTs (e.g. LV-ViT and Swin Transformer) are imported for the verification of the generalization of the winning tickets, and the respective methods are adjusted to adapt to the LTH experiments.

Why Conventional Winning Tickets Are Hard to Find on ViTs? – An Empirical Study

We try to identify the winning tickets in ViTs with the same settings of conventional LTH (Frankle and Carbin 2018). We implement the DeiT-Small (Touvron et al. 2021) in the experiments and obtain the sparse mask by pruning the pretrained model weights in layers of multi-head self-attention (MSA) and MLP blocks. Following the conventional LTH, we train the models from scratch with the proper or random weight initialization and applying the same sparse mask. The results of experiments implemented by DeiT-Small are shown in Table 1.

Although the model with proper initialization and low weight sparsity can match the accuracy of pretrained model, there is no accuracy difference (for the same weight sparsity) between the models trained from proper initialization (LTH) and random reinitialization (RR). According to Table 1, the performance of the LTH is lower than the performance of the RR with several different sparsity ratios, which means that the initializations are not as effective as in conventional LTH. Besides, we import the rewinding technique to improve the accuracy of the models trained from proper initialization (LTH). We achieve more than 2.4% in accuracy compared to the model trained from random reinitialization (RR) with 71% weight sparsity. However, there is still a 5% accuracy drop between the improved accuracy and the pretrain accuracy, which means the LTH is hard to match the pretrain accuracy even if the rewinding technique is used. Therefore, the subnetworks on weights that can be identified as winning tickets according to the conventional LTH are hard to find in ViTs.

Moreover, we also conduct the same experiments on a CNN where the original LTH is evaluated. We use a ResNet-20 (He et al. 2016) on CIFAR-10 (Krizhevsky 2009), and the results are showed in the Table 2. The results on ResNet-20 show that there is a small accuracy drop between the model trained from proper initialization (LTH) and pretrained model even at high weight sparsity (e.g. 90.1% vs. 90.3% Top-1 accuracy at 83.2% sparsity). The accuracy drop between the model trained from proper and random initialization is large (2.2% ~ 6.3% Top-1 accuracy difference under the same sparsity), which shows the identification of the conventional winning tickets. Furthermore, the accuracy is improved obviously when the rewinding technique is used in the experiments (e.g. 88.9% vs. 86.4% Top-1 accuracy at 95.6% sparsity).

These experimental results guide us to revisit the structure of ViTs, which might explain the difficulty of the identification of conventional winning tickets on ViTs. Unlike the CNNs convolutional kernel, which is an input-independent parameter of the static value. The matrix multiplication in the MSA module brings interaction between tokens, which means that ViT closely depends on the representation of the input. Due to the input-dependency characteristics of ViTs, the subnetwork produced by one sparse mask is difficult to generalize to the different input data. Therefore, we decide to identify winning tickets at the input level instead of model weights for ViTs.
| DeiT-Small | Weight Sparsity | LTH Acc. (%) | RR Acc. (%) | Acc. Diff. (%) |
|------------|----------------|--------------|-------------|---------------|
| None       | 0%             | 79.9         | 79.9        | 0             |
| MSA        | 12%            | 79.6         | 79.5        | 0.1           |
| MSA & MLP  | 44%            | 77.2         | 77.3        | -0.1          |
| MSA & MLP* | 44%            | 78.1         | 77.6        | 0.5           |
| MSA & MLP  | 53%            | 76.5         | 76.6        | -0.1          |
| MSA & MLP* | 71%            | 72.6         | 72.7        | -0.1          |
| MSA & MLP  | 71%            | 74.9         | 72.5        | 2.4           |

Table 1: Results of experiments following conventional LTH settings. The results in the table are based on DeiT-Small (Touvron et al. 2021) with ImageNet (Deng et al. 2009) dataset. The MSA indicates the multi-head self-attention layers and the MLP indicates the MLP blocks in the transformer encoders. LTH Acc. denotes the accuracy of the winning tickets (i.e., trained with a sparse mask from the appropriate initialization); RR Acc. denotes the accuracy of the model trained by the same mask but from another random reinitialization; Acc. Diff. denotes the accuracy difference between LTH Acc. and RR Acc.; and the * in the table denotes that the rewinding technique is used in experiments.

| ResNet-20 | Weight Sparsity | LTH Acc. (%) | RR Acc. (%) | Acc. Diff. (%) |
|-----------|----------------|--------------|-------------|---------------|
| None      | 0%             | 90.3         | 90.3        | 0             |
| All Conv layers | 83.2% | 90.1       | 87.9       | 2.2           |
| All Conv layers | 91.4% | 89.4       | 85.7       | 3.7           |
| All Conv layers* | 91.4% | 89.8       | 85.8       | 4.0           |
| All Conv layers | 94.5% | 87.5       | 83.9       | 3.6           |
| All Conv layers* | 94.5% | 89.4       | 83.9       | 5.5           |
| All Conv layers | 95.6% | 86.4       | 83.4       | 3.0           |
| All Conv layers* | 95.6% | 88.9       | 83.3       | 6.3           |

Table 2: The results in the table are based on all convolutional layers in ResNet-20 (He et al. 2016) on CIFAR-10 (Krizhevsky 2009).

**Patch Redundancy**

Transformer is first introduced to the computer vision research area by (Dosovitskiy et al. 2021). In this work, each image is first split into a sequence of fix-sized and non-overlapping patches, which are then projected to a sequence of tokens with embeddings before being fed into the transformer blocks. Each token is dependent on a unique patch of input image according to the positional embedding (Touvron et al. 2021). Furthermore, the tokens are independent of each other in the self-attention module of transformer blocks, which means the local information of each patch is not utilized. The reason why there exists redundancy in patches is that the input images are evenly segmented. Therefore, many patches may contain only background information and lack object information. Those patches contribute less to the classification task, so they can be removed without affecting the accuracy. Our choice of patch removal over weight pruning can be illustrated from two perspectives: 1) From the computation standpoint, it is the most efficient way to reduce model complexity. As shown in (Kong et al. 2021), under the same sparsity ratio, patch removal can reduce more overall computation than pruning channels and attention heads. 2) From a vision interpretability standpoint, weight pruning changes the model itself, which is input-independent. However, each image contains information in a different location and size. It is not possible to obtain one sparse model that is valid for all images. Patch removal can obtain different input sparsity for each image, which can significantly reduce the training and testing computation without reducing the model’s capability.

Previous works (Liang et al. 2022; Rao et al. 2021) use some additional optimization methods to identify and remove inattentive tokens at specific stages of the model. They argue that the tokens cannot be pruned at an early stage because: 1) The attention maps are unreliable for token removal. 2) Tokens may contain more global information, making it difficult to identify uninformative tokens. These difficulties have discouraged them from applying their methods at early stages, which creates a barrier to optimization. Therefore, it is important for us to identify the important tokens as early as possible in ViTs. Also, it is intuitive for us to maximize the reduction in computation by removing the inattentive image patches before feeding them into the first transformer block.

To achieve such a purpose, we insert a ticket selector to evaluate image patches and remove less important ones. Since the self-attention module in ViTs focuses on local information in early layers and global information in deeper layers, the ticket selector evaluate the importance of image patches according to the attention maps extracted from multiple depths of layers. Therefore, the selected image patches are the most attentive and informative ones, which can be defined as the winning tickets. More rigorous definition of winning tickets is provided in the definition section.

**Notations**

In this section, we create some notations and functions that are listed in Table 3 with detailed descriptions to illustrate our theory.

Consider a Transformer based network function \( f(\cdot) \) that
is denoted as \( f(x) \) when fed with the image \( x \). We define the following settings:

- **Pretraining:** We train the network \( f_B^R(\cdot) \) with the full image \( x \) for \( T \) epochs, arriving at \( f_B^R(\cdot) \).
- **Lottery Ticket:** We directly apply mask \( m \) to input image patches, resulting in \( x \odot m \) as winning tickets and the network function \( f_B^L(\cdot) \).
- **Random Choice:** We directly apply mask \( m' \) to input image patches, resulting in \( x \odot m' \) as the subset of randomly selected image patches and the network function \( f_R^L(\cdot) \).

### Definition of the Lottery Ticket Hypothesis for ViTs

According to our previous discussion, we have identified the feasibility of implementing the sparse ViTs by removing image patches at the input level instead of removing tokens at different stages inside the ViT architectures. In the following subsections, we propose the definition of the Lottery Ticket Hypothesis for Vision Transformers, and the principles of identifying the winning tickets for the ViTs. Additionally, for the mainstream ViT variants, we demonstrate the methodology for obtaining the corresponding winning tickets, which shows the universality and validity of our proposed definition.

**The Lottery Ticket Hypothesis for Vision Transformers.** For a ViT model and a non-trivial sparsity ratio, there exists a subset that contains the most attentive image patches that – when trained in isolation – can match\(^1\) the accuracy of the model that is trained with the full image, while showing a clear advantage in accuracy compared to the model that trained with a randomly selected subset with the same sparsity ratio for the same or less number of epochs.

In our definition of the LTH-ViTs, the subset of the input image patches is unique, in which each of the input images has a corresponding fixed mask that maintains a fixed topology of the dataset regardless of the sequence of the input images to the network. In order to achieve this, we make a reasonable analogy to the conventional LTH that uses the prior knowledge (i.e., the pretrained weights) to generate this unique mask. Based on the pretrained weights, we import the additional module \( p \) that can identify the importance of each token according to the attention map, and build the ticket selector as \( f_B^L(\cdot; p) \). More specifically, \( p \) the token identification module \( p \) is inserted before \( 4^{th}, 7^{th}, 10^{th} \) layers. \( p \) The module \( p \) sorts the tokens according to the [CLS] token \( \{\text{Dosovitskiy et al. 2021}\} \) which is allowed to interact with all the other tokens in attention map, and only keeps the top-k most informative tokens. \( \odot \) The attentive image patches can be indexed according to positional embeddings of the remaining tokens generated by the last module \( p \). After the attentive image patches are selected, we further explore the identification of the winning tickets in these selected image patches. We train the model \( f_B^L(\cdot) \) from scratch only with the subset of selected image patches \( x \odot m \) for \( T \) epochs and evaluate the performance of the model \( f_B^L(\cdot) \) being fed with the same subset \( x \odot m \) as \( f_B^L(x \odot m) \).

In the conventional LTH, the irreplaceable weight initialization value of the winning tickets makes up the unique winning property for the LTH. The performance of the subnetworks drops drastically when the initialization changes. In our proposed LTH-ViTs, the identified winning tickets have similar but not quite the same characteristic. The similarity is reflected in the fact that the winning tickets must use the same subset, and the difference is that this subset is within the input level. When the subset \( x \odot m \) is identified, the performance should surpass its randomly selected counterpart \( x \odot m' \) before calling it the winning ticket. Therefore, in our definition, it is necessary for us to explore the performance of the model trained with randomly selected image patches to validate the effectiveness and uniqueness of the winning tickets.

**The Identification of the Winning Tickets**

To make a valid identification of the winning ticket, three key evaluations should be performed: 1) the test accuracy of the ViT \( f_B^L(x \odot m) \), which is using the subset of attentive image patches \( x \odot m \) for each test image \( x \), 2) the test accuracy of ViT \( f_B^L(x \odot m) \), which is using identical test image subset \( x \odot m \) for testing, 3) the test accuracy of the ViT \( f_B^L(x) \) using full input image patches. By cross comparing the results of 1, 2 and 3, we can validate if the tickets are unique to be confidently identified as the winning ticket. When the test accuracy of \( f_B^L(x \odot m) \) can match the pretrained ViT \( f_B^L(x) \), and there exists a clear accuracy gap between \( f_B^L(x \odot m) \) and \( f_B^L(x \odot m) \), the subset of the input patches \( x \odot m \) can be considered as the winning ticket.

**Analogical similarity between conventional LTH and the proposed LTH-ViTs.** Both conventional LTH and LTH-ViTs attempt to identify the subsets of weights/inputs that can achieve similar accuracy to the model trained with full weights/inputs sets; 2) need to use prior knowledge (i.e. a pretrained model) to identify the winning tickets; 3) identify the winning tickets with the requirement of matching the accuracy of pretrained model; 4) import randomness into the subsets to validate the uniqueness and effectiveness of the winning tickets.

**Methodology**

Different variants of ViT are imported to demonstrate the effectiveness and generalization ability of our definition. The main architecture we use in this paper is DeiT \( \{\text{Touvron et al. 2021}\} \). To illustrate the integrity of our definition, we also explore the lottery ticket hypothesis for other ViT variants: Swin \( \{\text{Liu et al. 2021}\} \) and LV-ViT \( \{\text{Jiang et al. 2021}\} \), with some adjustments during the training and testing progress to match their model architecture. We use DeiT-Small as a ticket selector in experiments on those variants of ViTs.

**Vanilla ViT.** For the Lottery Ticket, we use the model DeiT-Small and DeiT-Tiny \( \{\text{Touvron et al. 2021}\} \) to validate the existence of winning tickets. We test our proposed LTH-ViTs definition using DeiT-Tiny and DeiT-Small. The overview of our method is shown in Figure

\(^1\) Accuracy drop between model trained with winning tickets and pretrain model is approximately within 0.5%. \( \{\text{Zhu and Gupta 2017}\}; \{\text{Frankle and Carbin 2018}\}; \{\text{Zhou et al. 2020}\} \)
the model to get \( f_{Ω}^{m}(x) \), which is the same as conventional LTH. ③ We incorporate the token identification module \( p \) into the model to build the ticket selector \( f_{Ω}^{m}(x; p) \). ④ We can use the ticket selector to produce the index of inattentive tokens for the mask \( m \) to build the winning tickets as \( x \odot m \). ⑤ We train the model only with the winning tickets for the same iterations as pretrained model. We evaluate the performance of the model trained with the full image \( x \) or winning tickets \( x \odot m \) to see if the test accuracy of the lottery ticket can match with the model trained and tested using full image patches. For the Random Choice, we replace the winning tickets with randomly selected patches with the same sparsity ratio during training, and the accuracy of the Random Choice is evaluated by using full image or identical test image patches that are generated by \( f_{Ω}^{m}(x; p) \) to ensure fairness.

**LV-ViT.** We use the LV-ViT-Small (Jiang et al. 2021) as the main architecture in experiments. According to the design of LV-ViT (Jiang et al. 2021), the token labeling technique is introduced to improve the model performance of the model. This method takes advantage of all tokens to compute the training loss in a dense manner instead of computing loss on the additional learnable class token (Jiang et al. 2021). To improve the recognition on the token level, each token in the LV-ViT is assigned with an additional label as location-specific supervision. Therefore, as Figure 2 (a) shows, we apply the same sparse mask on those labels as it on image patches to correct the computation of training loss, because the number of tokens is reduced when input image patches are applied with the sparse mask.

**Swin Transformer.** We mainly implement the Swin-Tiny (Liu et al. 2021b) in our experiments. We turn the pixel values of those inattentive patches to zero as the substitute for removal, and the pixels of the winning tickets produced by the ticket selector remain constant, as shown in Figure 2 (b). The purpose of using black occlusion rather than directly removing them is due to the introduction of shifted windows in the self-attention module of Swin Transformer (Liu et al. 2021b).

In this paper, all of the models are trained on the ImageNet (Deng et al. 2009) with approximately 1.2 million images in the training set and all of the results of accuracy are tested on the 50k images in the testing set. The image resolution in training and testing is 224 × 224, and the experiments are conducted with different sparsity ratios. As for the models of DeiT-Small and DeiT-Tiny (Touvron et al. 2021), we use the same architecture for both ticket selector and training model in the experiments. And as for the LV-ViT (Jiang et al. 2021) and Swin (Liu et al. 2021b), we introduce the DeiT-Small (Touvron et al. 2021) as the ticket selector to verify the generalization of the winning tickets. For different variants of models, we conduct a series of experiments with several sparsity ratios. For the training strategies and optimization methods, we follow the principles proposed by the original papers of DeiT (Touvron et al. 2021), Swin (Liu et al. 2021b) and LV-ViT (Jiang et al. 2021). For the testing process, all the models are tested in two different ways. We test the model with the winning tickets produced by the same ticket selector as it in training progress or with the full test images, which can demonstrate the identification of the winning tickets. All the experiments are conducted on the NVIDIA A100 with 8 GPUs.

**Main Results**

We report the main results using the models of DeiT-Small and DeiT-Tiny (Touvron et al. 2021) in Figure 3 with the top-1 accuracy vs. patch sparsity ratio curves plotted. The figure shows the results of two models with dense or sparse inference in four different colored lines varies with four different patch sparsity ratios. The Dense and Sparse inference mean that we evaluate the model with the full test image patches and the winning tickets respectively. And the LT (Lottery Ticket) and RC (Random Choice) denote that the model is trained with winning tickets and the subset of randomly selected patches respectively. It is obvious in the Figure 5 that the LT model colored with red and orange can achieve better accuracy than the RC model colored in green and blue. The accuracy drops between four lines become large with the increase in the sparsity ratio. Furthermore, according to the results of DeiT-Small, the LT model can match the accuracy of the pretrained model colored in grey with Sparse inference, because the accuracy drops between them are less

**Experiments and Results**

**Experimental Setups**

In this paper, all of the models are trained on the ImageNet (Deng et al. 2009) with approximately 1.2 million images in the training set and all of the results of accuracy are tested on the 50k images in the testing set. The image resolution in training and testing is 224 × 224, and the experiments are conducted with different sparsity ratios. As for the models of DeiT-Small and DeiT-Tiny (Touvron et al. 2021), we use the same architecture for both ticket selector and training model in the experiments. And as for the LV-ViT (Jiang et al. 2021) and Swin (Liu et al. 2021b), we introduce the DeiT-Small (Touvron et al. 2021) as the ticket selector to verify the generalization of the winning tickets. For different variants of models, we conduct a series of experiments with several sparsity ratios. For the training strategies and optimization methods, we follow the principles proposed by the original papers of DeiT (Touvron et al. 2021), Swin (Liu et al. 2021b) and LV-ViT (Jiang et al. 2021). For the testing process, all the models are tested in two different ways. We test the model with the winning tickets produced by the same ticket selector as it in training progress or with the full test images, which can demonstrate the identification of the winning tickets. All the experiments are conducted on the NVIDIA A100 with 8 GPUs.

**Main Results**

We report the main results using the models of DeiT-Small and DeiT-Tiny (Touvron et al. 2021) in Figure 3 with the top-1 accuracy vs. patch sparsity ratio curves plotted. The figure shows the results of two models with dense or sparse inference in four different colored lines varies with four different patch sparsity ratios. The Dense and Sparse inference mean that we evaluate the model with the full test image patches and the winning tickets respectively. And the LT (Lottery Ticket) and RC (Random Choice) denote that the model is trained with winning tickets and the subset of randomly selected patches respectively. It is obvious in the Figure 5 that the LT model colored with red and orange can achieve better accuracy than the RC model colored in green and blue. The accuracy drops between four lines become large with the increase in the sparsity ratio. Furthermore, according to the results of DeiT-Small, the LT model can match the accuracy of the pretrained model colored in grey with Sparse inference, because the accuracy drops between them are less
Figure 3: Top-1 accuracy vs. Patch sparsity on ImageNet. We show the lottery ticket experiments with DeiT Small and DeiT Tiny on ImageNet dataset and compare the experimental results with the pretrain accuracy. The LT and RC denote the model trained with Lottery Ticket and Random Choice respectively, and the Dense and Sparse indicates the model is tested with full image and winning tickets respectively.

than 0.5% when the sparsity is 27% and less than 1% even the sparsity is about 40%. As for the model of DeiT-Tiny, the accuracy drops between LT model and pretrained model at two smallest patch sparsity ratios are less than 1%, and the difference between LT model and RC model is noticeable as patch sparsity increases. Besides, we can find that the models trained and tested with winning tickets have the biggest advantage over others at the largest patch sparsity ratio as shown in the figure. Therefore, based on these observations, the identification of the winning tickets in normal ViTs can be demonstrated.

**Generalization Analysis**

Besides the DeiT model, we import LV-ViT [Jiang et al. 2021] and Swin Transformer [Liu et al. 2021b] to demonstrate the generalization of the winning tickets as shown in Table 4. The DeiT-Medium has a similar structure to DeiT-Small, while the dimension of embeddings in this model is changed into 576 and the number of heads is changed into 9. As for the Swin Transformer, we use the Swin-Tiny model with two different patch sparsity ratios to explore the identification of the winning tickets on Swin Transformer, and use all-zero values to occlude the inattentive patches to maintain the input size and window size during the training process. For LV-ViT, we train the LV-ViT-Small with token labels proposed by LV-ViT during the training process. All the accuracy of LT models can match the accuracy of the pretrained model at the given patch sparsity ratios and there are noticeable accuracy drops between LT models and RC models, which demonstrate the identification of the winning tickets. Thus, the generalization of the winning tickets through our definition of LTH for ViTs can be verified.

We also try to identify the winning tickets on CNNs with our proposed method, we use the ResNet-50 [He et al. 2016] to explore the identification of the winning tickets, and we use all-zero values to occlude the inattentive patches rather than cropping the image in the training progress. The experimental results show that CNNs is difficult to identify the new kind of winning tickets, because the LT models cannot match the pretrained accuracy and there is no accuracy drop between the LT models and RC models.

### Conclusion, Limitation, and Societal Impact

In this paper, we present the first Lottery Ticket Hypothesis for Vision Transformers (LTH-ViTs), which inherits the essential concepts and exhibits generalization to the conventional Lottery Ticket Hypothesis that is mainly discussed for CNNs. We conduct empirical studies to explain why the conventional lottery tickets are hard to find on ViTs, and propose our novel definition of LTH-ViTs and the principles for the identification of the winning tickets on the input of ViTs. We demonstrate that there exists a subset of input image patches (i.e., winning tickets of LTH-ViTs) that are capable of training a ViT and its multiple variants to achieve similar accuracy as the ones trained using full dataset. This paper mainly focuses on image classification tasks while there are ViTs variants designed for other tasks such as segmentation and detection, etc. We leave these tasks for our future work. The research in this paper is scientific in nature and we do not envision it generating any negative societal impact.
References

Bai, Y.; Wang, H.; TAO, Z.; Li, K.; and Fu, Y. 2022. Dual Lottery Ticket Hypothesis. In International Conference on Learning Representations.

Bao, H.; Dong, L.; and Wei, F. 2021. BEiT: BERT Pre-Training of Image Transformers. arXiv preprint arXiv:2106.08254.

Chavan, A.; Shen, Z.; Liu, Z.; Liu, Z.; Cheng, K.-T.; and Xing, E. 2022. Vision Transformer Slimming: Multi-Dimension Searching in Continuous Optimization Space. CVPR.

Chen, M.; Peng, H.; Fu, J.; and Ling, H. 2021a. Autoformer: Searching transformers for visual recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 12270–12280.

Chen, T.; Cheng, Y.; Gan, Z.; Yuan, L.; Zhang, L.; and Wang, Z. 2021b. Chasing sparsity in vision transformers: An end-to-end exploration. Advances in Neural Information Processing Systems, 34.

Chen, T.; Frankle, J.; Chang, S.; Liu, S.; Zhang, Y.; Wang, Z.; and Carbin, M. 2020. The Lottery Ticket Hypothesis for Pre-trained BERT Networks. arXiv:2007.12223.

Chen, T.; Zhang, Z.; Cheng, Y.; Awadallah, A.; and Wang, Z. 2022. The Principle of Diversity: Training Stronger Vision Transformers Calls for Reducing All Levels of Redundancy. arXiv preprint arXiv:2203.06345.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR.

El-Nouby, A.; Touvron, H.; Caron, M.; Bojanowski, P.; Douze, M.; Joulin, A.; Laptev, I.; Neverova, N.; Synnaeve, G.; Verbeek, J.; et al. 2021. XCiT: Cross-Covariance Image Transformers. arXiv preprint arXiv:2106.09681.

Frankle, J.; and Carbin, M. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. ICLR.

Frankle, J.; Dziugaite, G. K.; Roy, D.; and Carbin, M. 2020. Linear mode connectivity and the lottery ticket hypothesis. In International Conference on Machine Learning, 3259–3269. PMLR.

Gale, T.; Elsen, E.; and Hooker, S. 2019. The state of sparsity in deep neural networks. arXiv preprint arXiv:1902.09574.

Girish, S.; Maiya, S. R.; Gupta, K.; Chen, H.; Davis, L.; and Shrivastava, A. 2021. The Lottery Ticket Hypothesis for Object Recognition. CVPR.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In CVPR.

Hou, Z.; and Kung, S.-Y. 2021. Multi-Dimensional Model Compression of Vision Transformer. arXiv preprint arXiv:2201.00043.

Jiang, Z.; Hou, Q.; Yuan, L.; Zhou, D.; Shi, Y.; Jin, X.; Wang, A.; and Feng, J. 2021. All Tokens Matter: Token Labeling for Training Better Vision Transformers. arXiv preprint arXiv:2104.10858.

Kong, Z.; Dong, P.; Ma, X.; Meng, X.; Niu, W.; Sun, M.; Ren, B.; Qin, M.; Tang, H.; and Wang, Y. 2021. SPViT: Enabling Faster Vision Transformers via Soft Token Pruning. Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

Liang, Y.; Ge, C.; Tong, Z.; Song, Y.; Wang, J.; and Xie, P. 2022. Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations. In International Conference on Learning Representations.

Liu, Y.; Sangineto, E.; Bi, W.; Sebe, N.; Leprì, B.; and De Nadai, M. 2021a. Efficient Training of Visual Transformers with Small-SizeDatasets. arXiv preprint arXiv:2106.03746.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Ma, X.; Yuan, G.; Shen, X.; Chen, T.; Chen, X.; Chen, X.; Liu, N.; Qin, M.; Liu, S.; Wang, Z.; et al. 2021. Sanity Checks for Lottery Tickets: Does Your Winning Ticket Really Win the Jackpot? Advances in Neural Information Processing Systems, 34.

Mallya, A.; Davis, D.; and Lazebnik, S. 2018. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In Proceedings of the European Conference on Computer Vision (ECCV), 67–82.

Pan, B.; Panda, R.; Jiang, Y.; Wang, Z.; Feris, R.; and Oliva, A. 2021a. IA-RED: Interpretability-Aware Redundancy Reduction for Vision Transformers. Advances in Neural Information Processing Systems, 34.

Pan, Z.; Zhuang, B.; Liu, J.; He, H.; and Cai, J. 2021b. Scalable vision transformers with hierarchical pooling. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 377–386.

Prasanna, S.; Rogers, A.; and Rumshisky, A. 2020. When bert plays the lottery, all tickets are winning. EMNLP.

Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; and Hsieh, C.-J. 2021. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In Advances in Neural Information Processing Systems (NeurIPS).

Renda, A.; Frankle, J.; and Carbin, M. 2020. Comparing rewinding and fine-tuning in neural network pruning. ICLR.

Sai Prasanna, A. R., Anna Rogers. 2020. When BERT Plays the Lottery, All Tickets Are Winning. arXiv:2005.00561.

Shrey Desai, A. A., Hongyuan Zhan. 2019. Evaluating Lottery Tickets Under Distributional Shifts. arXiv:1910.12708.

Steiner, A.; Kolesnikov, A.; Zhai, X.; Wightman, R.; Uszkoreit, J.; and Beyer, L. 2021. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. arXiv preprint arXiv:2106.10270.
Tang, Y.; Han, K.; Wang, Y.; Xu, C.; Guo, J.; Xu, C.; and Tao, D. 2022. Patch slimming for efficient vision transformers. In CVPR.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jegou, H. 2021. Training data-efficient image transformers &amp; distillation through attention. In International Conference on Machine Learning, volume 139, 10347–10357.

Wang, P.; Wang, X.; Wang, F.; Lin, M.; Chang, S.; Xie, W.; Li, H.; and Jin, R. 2021. KVT: k-NN Attention for Boosting Vision Transformers. arXiv preprint arXiv:2106.00515.

Wu, K.; Peng, H.; Chen, M.; Fu, J.; and Chao, H. 2021. Re-thinking and improving relative position encoding for vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10033–10041.

Xu, Y.; Zhang, Z.; Zhang, M.; Sheng, K.; Li, K.; Dong, W.; Zhang, L.; Xu, C.; and Sun, X. 2021. Evo-ViT: Slow-Fast Token Evolution for Dynamic Vision Transformer. arXiv preprint arXiv:2108.01390.

Yu, F.; Huang, K.; Wang, M.; Cheng, Y.; Chu, W.; and Cui, L. 2022a. Width & Depth Pruning for Vision Transformers. AAAI.

Yu, S.; Chen, T.; Shen, J.; Yuan, H.; Tan, J.; Yang, S.; Liu, J.; and Wang, Z. 2022b. Unified Visual Transformer Compression. In International Conference on Learning Representations.

Yu, Z.; Fu, Y.; Li, S.; Li, C.; and Lin, Y. 2022c. MIA-Former: Efficient and Robust Vision Transformers via Multi-grained Input-Adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence.

Zhou, H.; Lan, J.; Liu, R.; and Yosinski, J. 2020. Deconstructing lottery tickets: Zeros, signs, and the supermask. Zhu, M.; and Gupta, S. 2017. To prune, or not to prune: exploring the efficacy of pruning for model compression. arXiv preprint arXiv:1710.01878.
Appendix

Comparisons with Existing Works

To demonstrate the effectiveness of winning tickets, we compare our methods with the existing methods in image classification task, including token pruning methods and weight pruning methods. In Table 6, we compare our method to several compression works with similar computation efforts on DeiT-Small (Touvron et al. 2021). According to Table 6, our work shows the existence of the model that can achieve the best accuracy compared to other ViT compression works. For those works that train ViTs from scratch, our work shows that the LT model trained with proposed winning tickets has clear advantage in accuracy compared to the works MIA-Former (Yu et al. 2022c) and HVT (Pan et al. 2021b). Besides, the LT model can achieve better accuracy than the work EViT (Liang et al. 2022) with smaller GMACs, and comparable accuracy with only half training epochs when compared to the work UVC (Yu et al. 2022b). Furthermore, we implement the same warmup strategy as EViT in our method. The warmup strategy allows the sparsity to gradually increase from 0 to the target value with a cosine schedule during the training process, which helps our method achieve better accuracy.

Conventional Winning Tickets in Swin Transformers

Besides implementing experiments on normal ViTs, we also try to identify the conventional winning tickets in Swin Transformers (Liu et al. 2021b) with the same settings as conventional LTH (Frankle and Carbin 2018). We implement the Swin-Tiny (Liu et al. 2021b) in the experiments, and obtain the sparse mask by pruning the pretrained model weights in layers of multi-head self-attention (MSA) and MLP blocks, which is similar to the experiments conducted on DeiT-Small described in Table 1. Following the conventional LTH, we train the models from scratch with the proper weight initialization (LTH) or random weight initialization (RR) and apply the same sparse mask. Also, we replace the sparse mask got from the pretrained model with a random mask (RM) to evaluate the network performance of different masks. The results are shown in Table 5. According to the results, there is no difference in accuracy between the model trained from different initializations. This indicates that there is no such unique subnetwork on weights that can be identified as a winning ticket according to the conventional LTH.

Visualization of Identified Winning Tickets in ViTs

We visualize the identified winning tickets on the input images of ViTs in Figure 4 (a). Moreover, we compare the sparsified input image and the progressively sparsified results at three different stages. We implement the DynamicViT (Rao et al. 2021) and EViT (Liang et al. 2022) with similar computation efforts as the winning tickets for comparison. The visualization of the DynamicViT is shown in Figure 4 (b) and the EViT is shown in Figure 4 (c). The figure shows that the remaining image patches of DynamicViT and EViT in the final stage are less than the winning tickets, and some attentive image patches are removed in the deep stage, which causes the accuracy drop compared to the winning tickets.

### Table 5: Results of experiments following conventional LTH settings on Swin-Tiny (Liu et al. 2021b). MSA indicates the multi-head self-attention layers and MLP indicates the MLP blocks in the transformer encoder. LTH Acc. denotes the accuracy of the winning tickets (i.e., trained from the appropriate initialization); RR Acc. denotes the accuracy of the model trained from another random reinitialization; and Acc. Diff. denotes the accuracy difference between LTH Acc. and RR Acc.

| Weights | Weight Sparsity | LTH Acc. (%) | RR Acc. (%) | Acc. Diff. (%) |
|---------|----------------|-------------|-------------|--------------|
| None    | 0%             | 81.2        | 81.2        | 0            |
| MSA & MLP | 42%           | 78.9        | 78.8        | 0.1          |
| MSA & MLP | 67%           | 74.2        | 74.8        | -0.6         |
| Model                        | MACs (G) | Epochs | Patch Sparsity | Token Sparsity | Weight Sparsity | Acc. (%) |
|------------------------------|----------|--------|----------------|----------------|-----------------|----------|
| DeiT-S                       | 4.6      | 300    | -              | -              | -               | 79.9     |
| DynamicViT-DeiT-S (Rao et al. 2021) | 2.2      | 30     | ✓              | ✓              | ✓               | 77.5     |
| WDPruning-DeiT-S (Yu et al. 2022a) | 3.1      | 100    | ×              | ×              | ✓               | 78.6     |
| WDPruning-DeiT-S (Yu et al. 2022a) | 2.6      | 100    | ×              | ×              | ✓               | 78.4     |
| ViT-Slim<sub>P_S</sub>-DeiT-S (Chavan et al. 2022) | 2.3      | 350    | ×              | ×              | ✓               | 77.9     |
| EViT-DeiT-S (Liang et al. 2022) | 2.3      | 30     | ×              | ✓              | ✓               | 78.5     |
| EViT-DeiT-S (Liang et al. 2022) | 2.3      | 300    | ×              | ✓              | ×               | 78.5     |
| MIA-Former-DeiT-S (Yu et al. 2022c) | 2.0      | -      | ×              | ✓              | ✓               | 78.1     |
| HVT-DeiT-S (Pan et al. 2021b) | 2.4      | 300    | ×              | ✓              | ✓               | 78.0     |
| UVC-DeiT-S (Yu et al. 2022b) | 2.3      | 600    | ×              | ✓              | ✓               | 78.8     |
| Ours-DeiT-S                  | 2.2      | 300    | ✓              | ×              | ×               | 78.7     |
| Ours-DeiT-S*                 | 2.2      | 300    | ✓              | ×              | ×               | 78.9     |

Table 6: Comparisons with the existing works with comparable MACs on ImageNet. The result with * means that the same warmup strategy as EViT is used during the training progress.

Figure 4: Visualization of the sparsified input image and progressively sparsified results. (a) Identified winning tickets, the subset of the most attentive image patches. (b) DynamicViT (Rao et al. 2021), removing tokens at three different stages with additional token predictors. (c) EViT (Liang et al. 2022), removing tokens at three different stages according to the importance of tokens.