The Lottery Ticket Hypothesis

Finding Sparse, Trainable Neural Networks

Matthia Sabatelli[1]

March 5, 2020

[1]Montefiore Institute, Department of Electrical Engineering and Computer Science, Université de Liège, Belgium

Presentation outline

1. The Lottery Ticket Hypothesis (LTH)
2. Towards a Better Understanding of the LTH
3. On the Generalization Properties of Lottery-Winners

An overview of popular Deep Learning Models

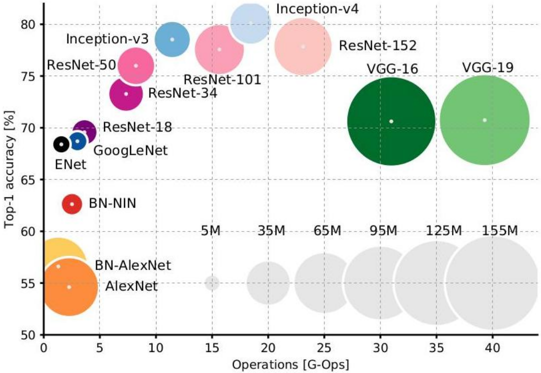


Figure: Image taken from Canziani A. et al. An Analysis of Deep Neural Network Models for Practical Applications.

The Lottery Ticket Hypothesis (LTH)

The Lottery Ticket Hypothesis (LTH)

Network Pruning

To reduce the extent of [a neural network] by removing its superfluous and unwanted parts [its weights].

The Lottery Ticket Hypothesis (LTH)

Network Pruning

To reduce the extent of [a neural network] by removing its superfluous and unwanted parts [its weights].

- Reduce computational and memory requirements

The Lottery Ticket Hypothesis (LTH)

Network Pruning

To reduce the extent of [a neural network] by removing its superfluous and unwanted parts [its weights].

- Reduce computational and memory requirements
- Fit large models on e.g. smartphones

The Lottery Ticket Hypothesis (LTH)

Network Pruning

To reduce the extent of [a neural network] by removing its superfluous and unwanted parts [its weights].

- Reduce computational and memory requirements
- Fit large models on e.g. smartphones
- Faster and more efficient inference

The Lottery Ticket Hypothesis (LTH)

Network Pruning

To reduce the extent of [a neural network] by removing its superfluous and unwanted parts [its weights].

- Reduce computational and memory requirements
- Fit large models on e.g. smartphones
- Faster and more efficient inference
- (In some cases) lead to better performance

The Lottery Ticket Hypothesis (LTH)

- An idea which is **not new** ...

598 Le Cun, Denker and Solla

Optimal Brain Damage

Yann Le Cun, John S. Denker and Sara A. Solla
AT&T Bell Laboratories, Holmdel, N. J. 07733

ABSTRACT

We have used information-theoretic ideas to derive a class of practical and nearly optimal schemes for adapting the size of a neural network. By removing unimportant weights from a network, several improvements can be expected: better generalization, fewer training examples required, and improved speed of learning and/or classification. The basic idea is to use second-derivative information to make a tradeoff between network complexity and training set error. Experiments confirm the usefulness of the methods on a real-world application.

The Lottery Ticket Hypothesis (LTH)

- The benefits of pruning were already known

The Lottery Ticket Hypothesis (LTH)

- The benefits of pruning were already known
- If a network can be reduced in size, why don't we train smaller architectures to begin with?

The Lottery Ticket Hypothesis (LTH)

- The benefits of pruning were already known
- If a network can be reduced in size, why don't we train smaller architectures to begin with?

Because it was considered impossible

The Lottery Ticket Hypothesis (LTH)

- The benefits of pruning were already known
- If a network can be reduced in size, why don't we train smaller architectures to begin with?

Because it **was** considered impossible

- "Training a pruned-model from scratch performs worst than fine-tuning a pruned model which has already gone through training before."

The Lottery Ticket Hypothesis (LTH)

- The benefits of pruning were already known
- If a network can be reduced in size, why don't we train smaller architectures to begin with?

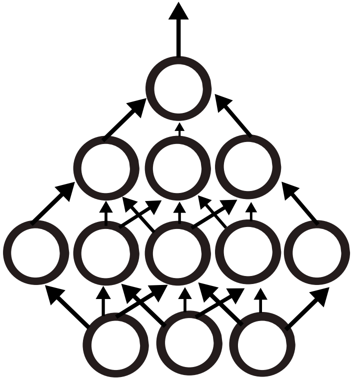
Because it was considered impossible

- "Training a pruned-model from scratch performs worst than fine-tuning a pruned model which has already gone through training before."
- "It is better to retain the weights from the initial training phase than it is to re-initialize the pruned model."

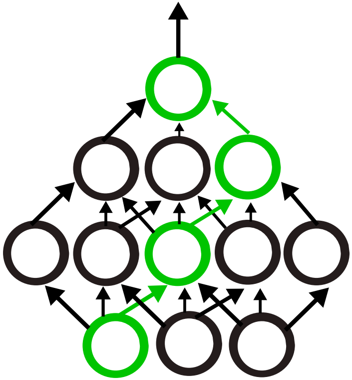
The Lottery Ticket Hypothesis (LTH)

The Lottery Ticket Hypothesis: *A randomly-initialized dense neural network contains a subnetwork that is initialized such that -when trained in isolation- it can match the test accuracy of the original network after training for at most the same number of iterations.*

The Lottery Ticket Hypothesis (LTH)



The Lottery Ticket Hypothesis (LTH)



The Lottery Ticket Hypothesis (LTH)

How do we **find** the subnetworks that are the winners of the LTH?

1. Randomly initialize a network $f(x; \theta_0)$ where $\theta_0 \sim \mathcal{D}_\theta$

The Lottery Ticket Hypothesis (LTH)

How do we **find** the subnetworks that are the winners of the LTH?

1. Randomly initialize a network $f(x; \theta_0)$ where $\theta_0 \sim \mathcal{D}_\theta$
2. Train the network for j iterations

The Lottery Ticket Hypothesis (LTH)

How do we **find** the subnetworks that are the winners of the LTH?

1. Randomly initialize a network $f(x; \theta_0)$ where $\theta_0 \sim \mathcal{D}_\theta$
2. Train the network for j iterations
3. Prune $p\%$ of the parameters in θ_j, creating a mask m

The Lottery Ticket Hypothesis (LTH)

How do we **find** the subnetworks that are the winners of the LTH?

1. Randomly initialize a network $f(x; \theta_0)$ where $\theta_0 \sim \mathcal{D}_\theta$
2. Train the network for j iterations
3. Prune $p\%$ of the parameters in θ_j, creating a mask m
4. Reset the remaining parameters to their values at θ_0 (**and not at θ_j!**), creating a winning-ticket $f(x; m \odot \theta_0)$

The Lottery Ticket Hypothesis (LTH)

Winning-Tickets in Fully-Connected Networks

- Consider a MLP which gets trained on the MNIST dataset

The Lottery Ticket Hypothesis (LTH)

Winning-Tickets in Fully-Connected Networks

- Consider a MLP which gets trained on the MNIST dataset
- Weights get pruned based on their **magnitude**

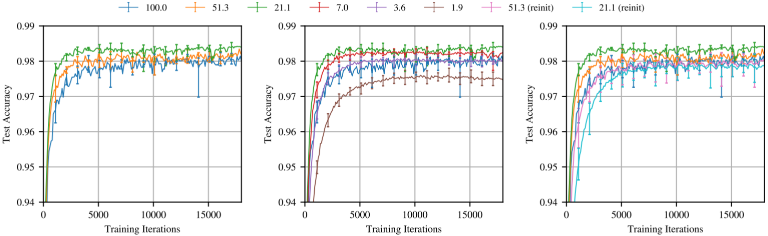
The Lottery Ticket Hypothesis (LTH)

Winning-Tickets in Fully-Connected Networks

- Consider a MLP which gets trained on the MNIST dataset
- Weights get pruned based on their **magnitude**
- Winning tickets get compared to sparse models which get randomly reinitialized $\theta' \sim \mathcal{D}_\theta$ and are therefore not the winners of the lottery anymore

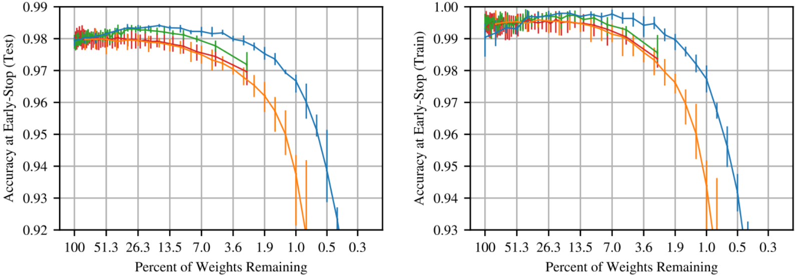
The Lottery Ticket Hypothesis (LTH)

- An overall performance of sparse models winners of the LTH



The Lottery Ticket Hypothesis (LTH)

- A comparison with randomly initialized sparse models



The Lottery Ticket Hypothesis (LTH)

Winning-Tickets in Convolutional Networks

- Consider a VGG-like architecture which gets trained on the CIFAR-10 dataset

The Lottery Ticket Hypothesis (LTH)

Winning-Tickets in Convolutional Networks

- Consider a VGG-like architecture which gets trained on the CIFAR-10 dataset
- Comparison to the **Dropout** regularization technique

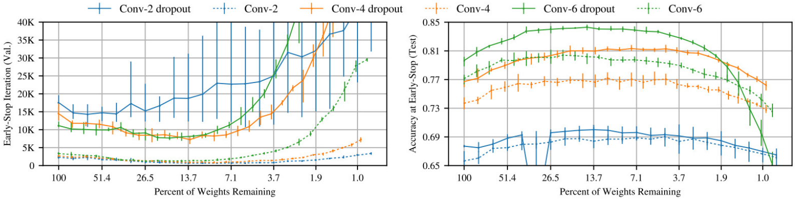
The Lottery Ticket Hypothesis (LTH)

Winning-Tickets in Convolutional Networks

- Consider a VGG-like architecture which gets trained on the CIFAR-10 dataset
- Comparison to the **Dropout** regularization technique
- Extension to more popular architectures such a ResNet-50

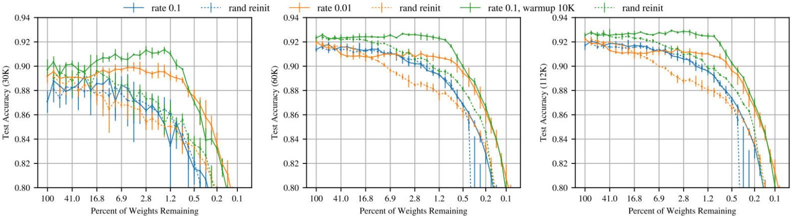
The Lottery Ticket Hypothesis (LTH)

- Convolutional Networks and Dropout regularization



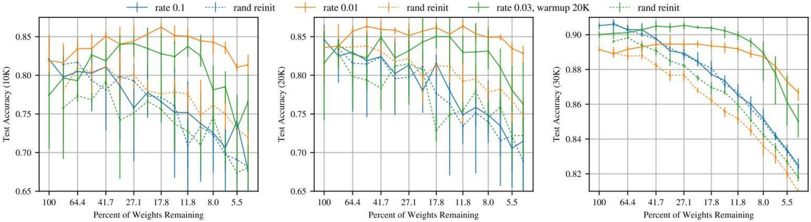
The Lottery Ticket Hypothesis (LTH)

- Deeper Convolutional Networks (VGG-19) on CIFAR-10



The Lottery Ticket Hypothesis (LTH)

- Deeper Convolutional Networks (ResNet-18) on CIFAR-10



The Lottery Ticket Hypothesis (LTH)

- Results are very consistent among all the tested experimental setups

The Lottery Ticket Hypothesis (LTH)

- Results are very consistent among all the tested experimental setups
- Additional extensions include different optimization algorithms, regularizers, etc ...

The Lottery Ticket Hypothesis (LTH)

- Results are very consistent among all the tested experimental setups
- Additional extensions include different optimization algorithms, regularizers, etc ...
- Show that pruned models can be trained **from scratch**

The Lottery Ticket Hypothesis (LTH)

- Results are very consistent among all the tested experimental setups
- Additional extensions include different optimization algorithms, regularizers, etc ...
- Show that pruned models can be trained **from scratch**
- Albeit this seems to be harder to achieve when CNNs are considered

The Lottery Ticket Hypothesis (LTH)

- What makes some weights **so special** to be the winners of the LTH?
    1. Some of the initial weights in θ_0 are already close to the weights we would obtain after gradient descent

The Lottery Ticket Hypothesis (LTH)

- What makes some weights **so special** to be the winners of the LTH?
  1. Some of the initial weights in θ_0 are already close to the weights we would obtain after gradient descent
  2. Winning tickets are a compromise between large overparametrized models and too small ones

The Lottery Ticket Hypothesis (LTH)

- What makes some weights **so special** to be the winners of the LTH?
    1. Some of the initial weights in θ_0 are already close to the weights we would obtain after gradient descent
    2. Winning tickets are a compromise between large overparametrized models and too small ones
    3. If understood this could lead to better weights initialization strategies (we're far from this)

The Lottery Ticket Hypothesis (LTH)

- What are the **limitations** of the paper and of the proposed approach?
  1. Identifying the weights which are the winners of the LTH is computationally very expensive

The Lottery Ticket Hypothesis (LTH)

- What are the **limitations** of the paper and of the proposed approach?
    1. Identifying the weights which are the winners of the LTH is computationally very expensive
    2. There is no *a-priori* way for identifying which θ_0 will be the winners of the lottery

The Lottery Ticket Hypothesis (LTH)

- What are the **limitations** of the paper and of the proposed approach?
  1. Identifying the weights which are the winners of the LTH is computationally very expensive
  2. There is no *a-priori* way for identifying which θ_0 will be the winners of the lottery
  3. The paper only considers 2 classification problems on 2 relatively simple datasets

The Lottery Ticket Hypothesis (LTH)

- What are the **limitations** of the paper and of the proposed approach?
    1. Identifying the weights which are the winners of the LTH is computationally very expensive
    2. There is no *a-priori* way for identifying which θ_0 will be the winners of the lottery
    3. The paper only considers 2 classification problems on 2 relatively simple datasets
    4. Presents this new deep-learning phenomenon but does not give any insights on why this happens

Towards a Better Understanding of the LTH

Towards a Better Understanding of the LTH

- After the original paper was introduced it appeared that finding lottery-winners in CNNs was much harder than expected!

Towards a Better Understanding of the LTH

- After the original paper was introduced it appeared that finding lottery-winners in CNNs was much harder than expected!
- New formalization of the LTH phenomenon

Towards a Better Understanding of the LTH

- After the original paper was introduced it appeared that finding lottery-winners in CNNs was much harder than expected!
- New formalization of the LTH phenomenon
- Researchers started to focus whether winning-initializations could be transferred among domains

Towards a Better Understanding of the LTH

Let's take a look again at the way winning tickets should be found...

1. Randomly initialize a network $f(x; \theta_0)$ where $\theta_0 \sim \mathcal{D}_\theta$
2. Train the network for j iterations
3. Prune $p\%$ of the parameters in θ_j, creating a mask m
4. Reset the remaining parameters to their values at θ_0 (**and not at θ_j!**), creating a winning-ticket $f(x; m \odot \theta_0)$

Towards a Better Understanding of the LTH

Let's take a look again at the way winning tickets should be found...

1. Randomly initialize a network $f(x; \theta_0)$ where $\theta_0 \sim \mathcal{D}_\theta$
2. Train the network for j iterations
3. Prune $p\%$ of the parameters in θ_j, creating a mask m
4. Reset the remaining parameters to their values at θ_0 (**and not at θ_j!**), creating a winning-ticket $f(x; m \odot \theta_0)$

Towards a Better Understanding of the LTH

To be be a successful lottery-ticket hunter ...

- In order to find winning-tickets in CNNs **late-resetting** needs to be used.
- Train the original θ parameters for k iterations, without going through any pruning
- Then start the usual procedure for finding winning-tickets
- k is an hyperparameter that needs to be tuned, therefore not leading to a fully trained model!

Towards a Better Understanding of the LTH

If we take a look at the new LTH scheme ...

1. Randomly initialize a network $f(x;\theta_0)$ where $\theta_0 \sim \mathcal{D}_\theta$
2. Train the network for k iterations
3. Prune $p\%$ of the parameters in θ_k, creating a mask m
4. Reset the remaining parameters to their values at θ_k, creating a winning-ticket $f(x; m \odot \theta_0)$

Towards a Better Understanding of the LTH

- Once this **rewinding trick** was introduced finding the winners of the lottery became much easier
- The presence of lottery-winners has been found in a large set of architectures trained on even larger datasets
- Therefore characterizing the phenomenon better (when do these weights appear?)
- Not providing an answer to **why** these weights appear at the early stages of training

Towards a Better Understanding of the LTH

Now that finding winning-tickets should be (relatively) easy, what are the **generalization** properties of lottery-winners

- Are winning-tickets dataset dependent?

Towards a Better Understanding of the LTH

Now that finding winning-tickets should be (relatively) easy, what are the **generalization** properties of lottery-winners

- Are winning-tickets dataset dependent?
- Are winning-tickets optimizer dependent?

Towards a Better Understanding of the LTH

Now that finding winning-tickets should be (relatively) easy, what are the **generalization** properties of lottery-winners

- Are winning-tickets dataset dependent?
- Are winning-tickets optimizer dependent?
- Do they contain inductive-biases which are independent from the training scenario these tickets have been found on?

Towards a Better Understanding of the LTH

Now that finding winning-tickets should be (relatively) easy, what are the **generalization** properties of lottery-winners

- Are winning-tickets dataset dependent?
- Are winning-tickets optimizer dependent?
- Do they contain inductive-biases which are independent from the training scenario these tickets have been found on?

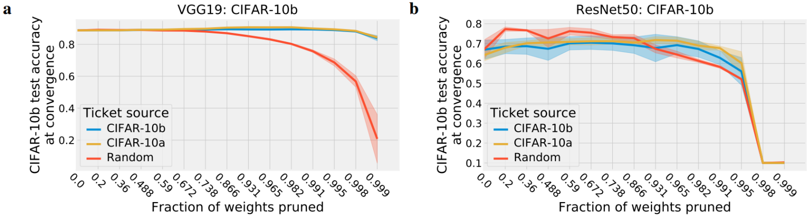
Why should we care?

Once answered, all these questions could give insights about **what** makes some weights so special to be the winners of the initialization lottery!

On the Generalization Properties of the LTH

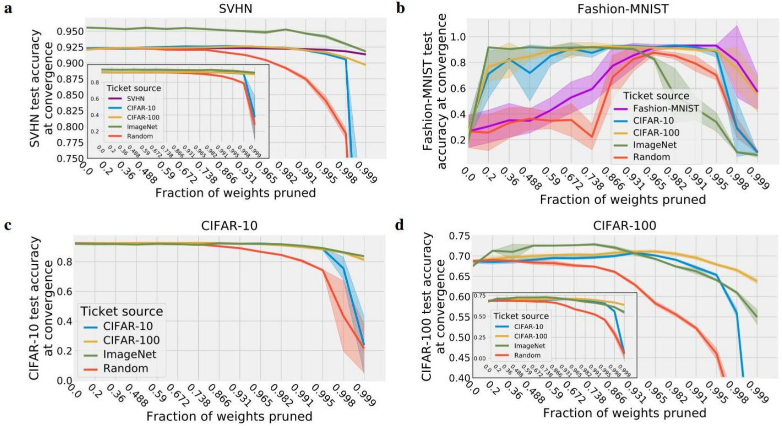
On the Generalization Properties of the LTH

- Do winning tickets generalize within the **same** data distribution?



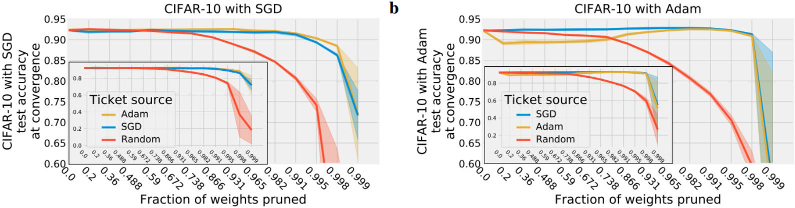
On the Generalization Properties of the LTH

- Do winning tickets generalize between **several** photorealistic data distributions?



On the Generalization Properties of the LTH

- Do winning tickets generalize across **optimizers**?



On the Generalization Properties of the LTH

- It does seem that winning-initializations are very **general** and that they are independent from a considered experimental setup
- **However** although different datasets have been used, all images come from the same source domain (natural images)
- Do lottery winners generalize to very different distributions?
- Can we really find one ticket to rule them all?! and therefore understand the LT phenomenon even more?

On the Generalization Properties of the LTH

This is what I explore in my latest paper ...

- Do winning tickets contain inductive biases which are domain independent?

On the Generalization Properties of the LTH

This is what I explore in my latest paper ...

- Do winning tickets contain inductive biases which are domain independent?
- Is it worth finding lottery winners?

On the Generalization Properties of the LTH

This is what I explore in my latest paper ...

- Do winning tickets contain inductive biases which are domain independent?
- Is it worth finding lottery winners?
    1. Finding lottery winners is computationally very expensive

On the Generalization Properties of the LTH

This is what I explore in my latest paper ...

- Do winning tickets contain inductive biases which are domain independent?
- Is it worth finding lottery winners?
  1. Finding lottery winners is computationally very expensive
  2. Although the benefits of the found tickets are clear, do they compensate for all the computing time that is required for finding them?

On the Generalization Properties of the LTH

This is what I explore in my latest paper ...

- Do winning tickets contain inductive biases which are domain independent?
- Is it worth finding lottery winners?
  1. Finding lottery winners is computationally very expensive
  2. Although the benefits of the found tickets are clear, do they compensate for all the computing time that is required for finding them?
- Can this help us characterize the LTH even better?

On the Generalization Properties of the LTH

An additional benefit that so far has not been considered ...

Abstract

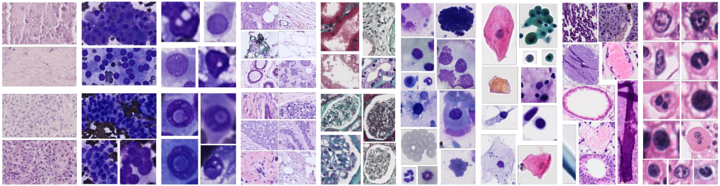
Neural network pruning techniques can reduce the parameter counts of trained networks by over 90%, decreasing storage requirements and improving computational performance of inference without compromising accuracy. However, contemporary experience is that the sparse architectures produced by pruning are difficult to train from the start, which would similarly improve training performance.

We find that a standard pruning technique naturally uncovers subnetworks whose initializations made them capable of training effectively. Based on these results, we articulate the *lottery ticket hypothesis*: dense, randomly-initialized, feed-forward networks contain subnetworks (*winning tickets*) that—when trained in isolation—reach test accuracy comparable to the original network in a similar number of iterations. The winning tickets we find have won the initialization lottery: their connections have initial weights that make training particularly effective.

We present an algorithm to identify winning tickets and a series of experiments that support the lottery ticket hypothesis and the importance of these fortuitous initializations. We consistently find winning tickets that are less than 10-20% of the size of several fully-connected and convolutional feed-forward architectures for MNIST and CIFAR10. Above this size, the winning tickets that we find learn faster than the original network and reach higher test accuracy.

On the Generalization Properties of the LTH

- In training conditions where data is scarce and therefore it results hard to train large networks identifying the lottery winners can be particularly important
- Can we avoid finding new lottery winners for each dataset and simply reuse already pruned models?

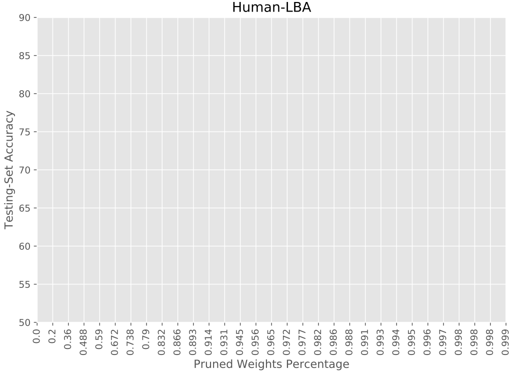


(a) Necrosis (b) ProliferativePattern (c) CellInclusion (d) Breast (e) Glomeruli (f) MouseLba (g) HumanLba (h) Lung (i) BoneMarrow

Figure: Image taken from Mormont R. et al. Comparison of deep transfer learning strategies for digital pathology.

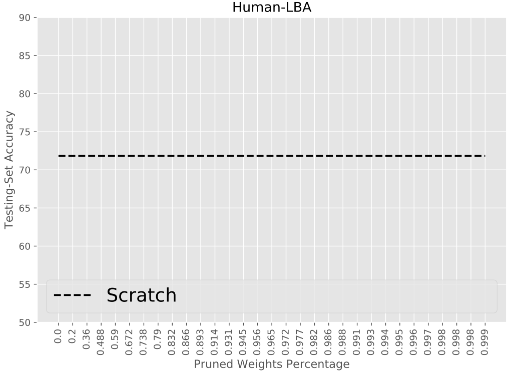
On the Generalization Properties of the LTH

- Is it worth it seeking for lottery winners?



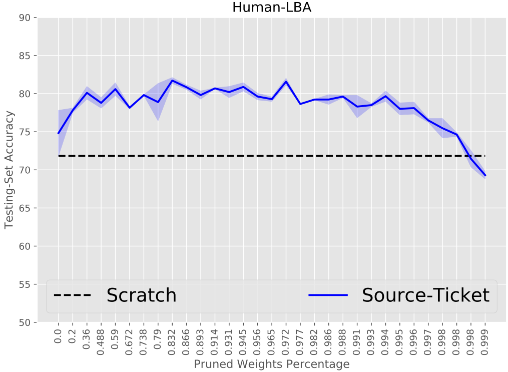
On the Generalization Properties of the LTH

- Is it worth it seeking for lottery winners?



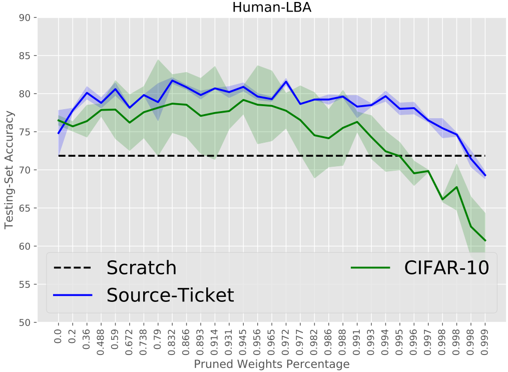
On the Generalization Properties of the LTH

- Is it worth it seeking for lottery winners?



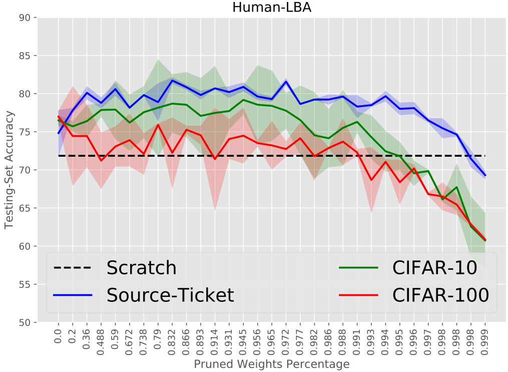
On the Generalization Properties of the LTH

- Is it worth it seeking for lottery winners?



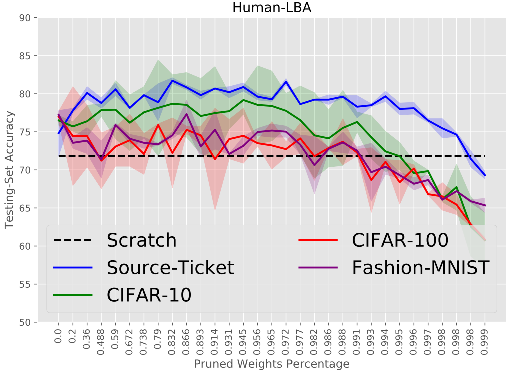
On the Generalization Properties of the LTH

- Is it worth it seeking for lottery winners?



On the Generalization Properties of the LTH

- Is it worth it seeking for lottery winners?

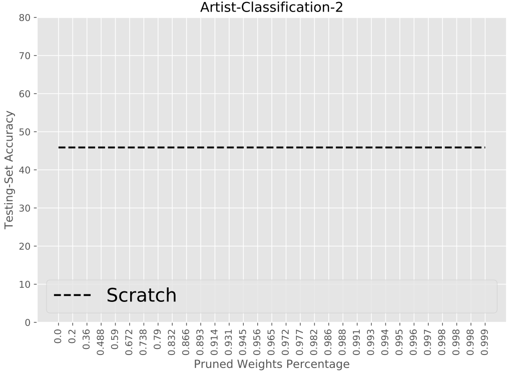


On the Generalization Properties of the LTH

- Do winning initializations transfer as well as claimed before?

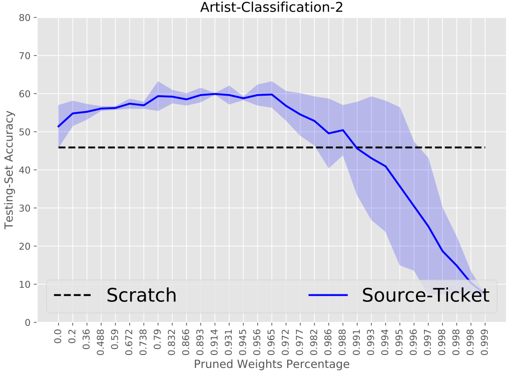
On the Generalization Properties of the LTH

- Do winning initializations transfer as well as claimed before?



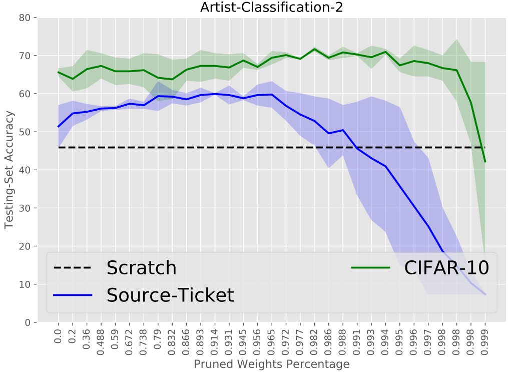
On the Generalization Properties of the LTH

- Do winning initializations transfer as well as claimed before?



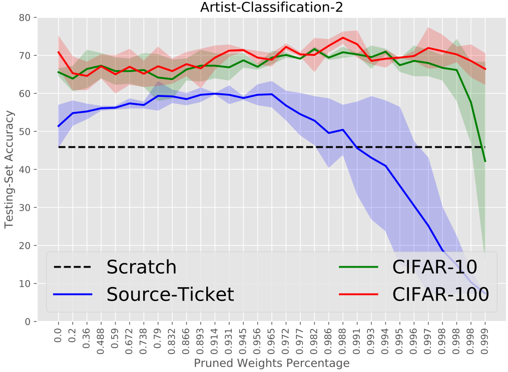
On the Generalization Properties of the LTH

- Do winning initializations transfer as well as claimed before?



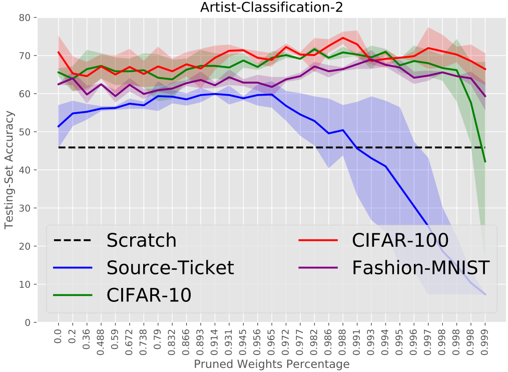
On the Generalization Properties of the LTH

- Do winning initializations transfer as well as claimed before?



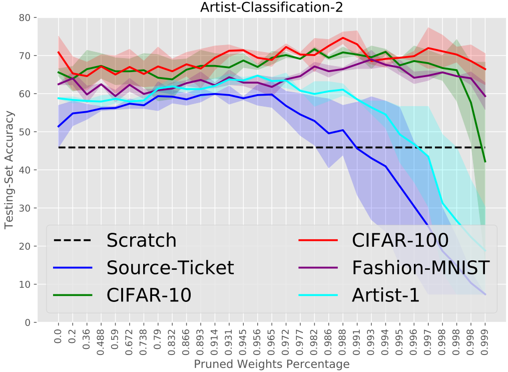
On the Generalization Properties of the LTH

- Do winning initializations transfer as well as claimed before?



On the Generalization Properties of the LTH

- Do winning initializations transfer as well as claimed before?



On the Generalization Properties of the LTH

Some final considerations about the LTH

- It is beneficial to look for lottery-winners in order to maximize the performance of deep neural networks

On the Generalization Properties of the LTH

Some final considerations about the LTH

- It is beneficial to look for lottery-winners in order to maximize the performance of deep neural networks
- Especially when training data is scarce, small models winners of the lottery **always** perform better than their larger overparametrized counterparts

On the Generalization Properties of the LTH

Some final considerations about the LTH

- It is beneficial to look for lottery-winners in order to maximize the performance of deep neural networks
- Especially when training data is scarce, small models winners of the lottery **always** perform better than their larger overparametrized counterparts
- There are limitations to the transferability of winning initializations between domains

On the Generalization Properties of the LTH

Some final considerations about the LTH

- It is beneficial to look for lottery-winners in order to maximize the performance of deep neural networks
- Especially when training data is scarce, small models winners of the lottery **always** perform better than their larger overparametrized counterparts
- There are limitations to the transferability of winning initializations between domains

Last but not least: what makes some weights so special?!

We still have no idea!

References and Credits

- Frankle, Jonathan, and Michael Carbin. "The lottery ticket hypothesis: Finding sparse, trainable neural networks." arXiv preprint arXiv:1803.03635 (2018).
- Frankle, Jonathan, et al. "Stabilizing the Lottery Ticket Hypothesis." arXiv, page.
- Yu, Haonan, et al. "Playing the lottery with rewards and multiple languages: lottery tickets in RL and NLP." arXiv preprint arXiv:1906.02768 (2019).

References and Credits

- Morcos, Ari, et al. "One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers." Advances in Neural Information Processing Systems. 2019.
- Li, Hao, et al. "Pruning filters for efficient convnets." arXiv preprint arXiv:1608.08710 (2016).
- Frankle, Jonathan, et al. "The Early Phase of Neural Network Training" arXiv preprint arXiv:2002.10365 (2020).