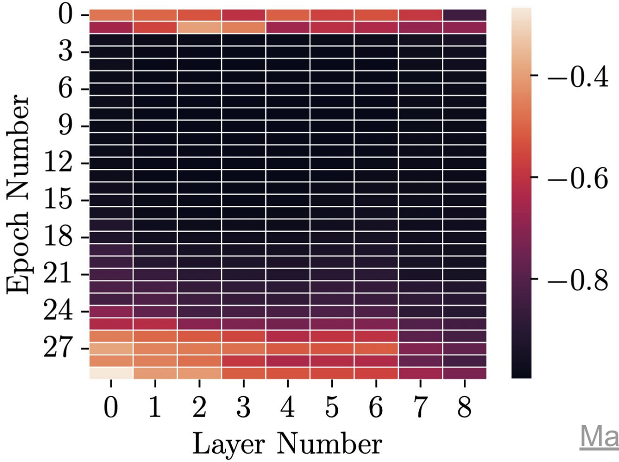
Neural Network Memorization

Sergey Sedov

Is memorization that bad?

- On the corrupt or noisy data everything is pretty obvious - we want our models to be independent from it and generalize well on the rest.
- However, there is a fraction of atypical examples - outliers, that may reveal in a similar way. We still want to surpass memorization, but it may be necessary to memorize them at the intermediate steps. For instance, gradients of regular and atypical examples could behave similarly to the mislabeled ones:

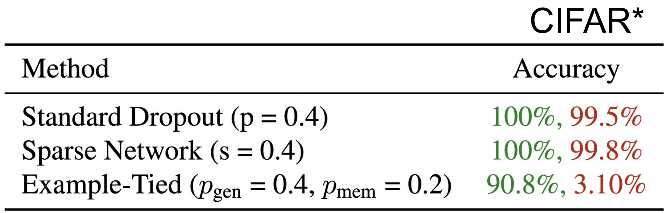
Figure 2. Cosine similarity between the average gradients of clean and mislabeled examples per layer, per epoch for ResNet9 model on the CIFAR10 dataset with 10% label noise. The memorization of mislabeled examples happens between epochs 10–30 (Figure 3).



Maini et al., 2023

Is memorization that bad?

- On the corrupt or noisy data everything is pretty obvious - we want our models to be independent from it and generalize well on the rest.
- However, there is a fraction of atypical examples - outliers, that may reveal in a similar way. We still want to surpass memorization, but it may be necessary to memorize them at the intermediate steps. And the example-tied dropout harms performance on outliers too:



	CIFAR*
Method	Accuracy
Standard Dropout (p = 0.4)	100%, 99.5%
Sparse Network (s = 0.4)	100%, 99.8%
Example-Tied ($p_{\text{gen}} = 0.4$, $p_{\text{mem}} = 0.2$)	90.8%, 3.10%

Table 3. Comparison of Accuracy on clean, noisy training examples after training with various methods that activate only a small fraction of the network during training time.

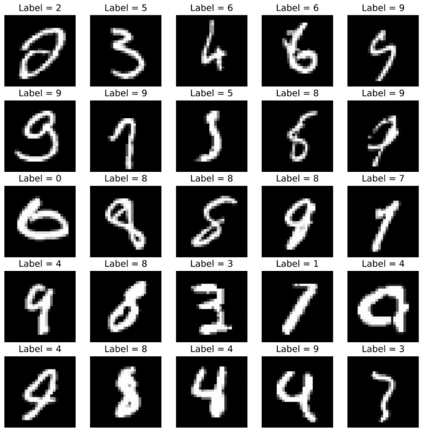
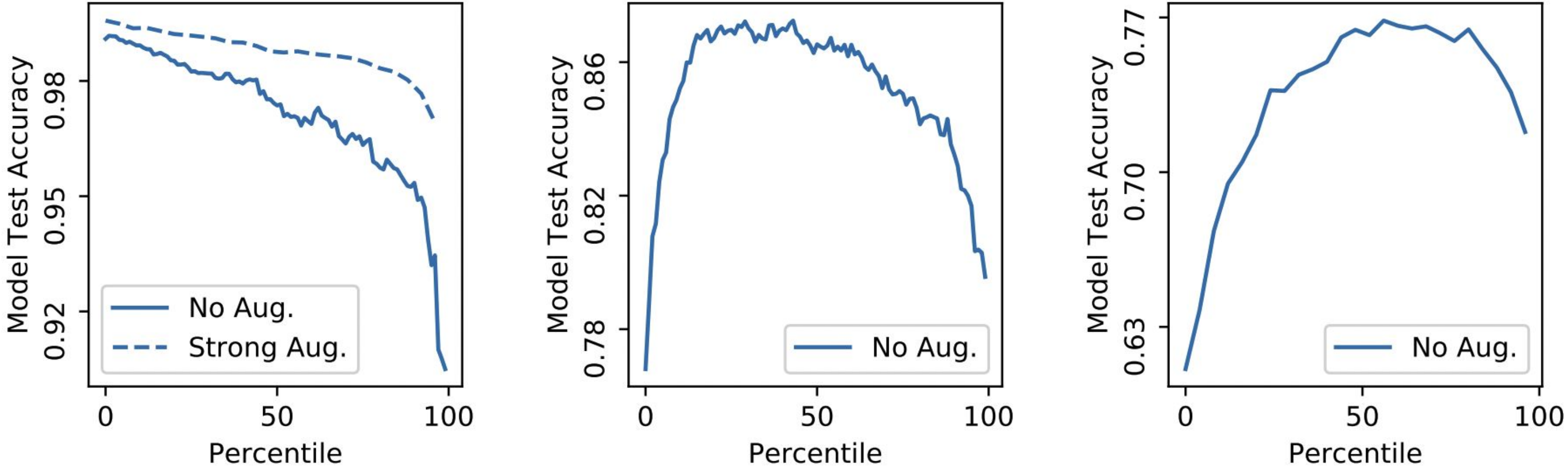


Figure 7. Most of the clean examples that are forgotten when dropping out the neurons responsible for memorization in the case of Example-tied dropout were either mislabeled or inherently ambiguous and unique requiring memorization for correct classification.

Maini et al., 2023

Do we need outliers?

Carlini et al., 2019



(a) MNIST

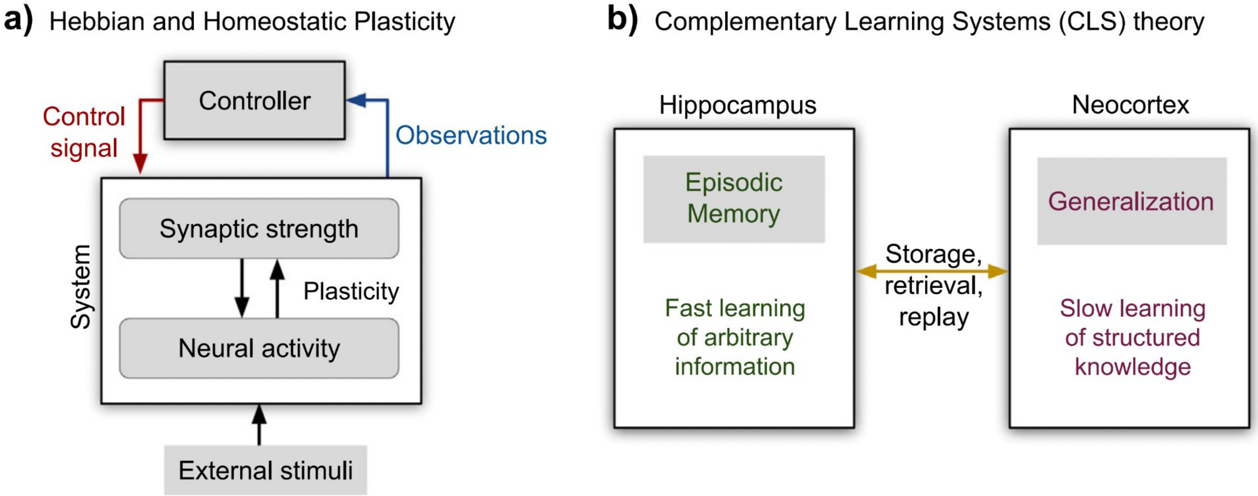
(b) Fashion-MNIST

(c) CIFAR-10

Figure 6: Final test accuracy of a model after trained on $5,000$ training examples consecutively ranked by the adv metric (so that training on the least representative examples are that the 0th percentile, and the most representative at the 100th percentile). See text for full details. Given only $5,000$ training examples, on MNIST (subplot (a)) training on the outliers is always better, however for Fashion-MNIST (b) and CIFAR-10 (c) it is preferable to train using neither the most nor least well-represented examples, but those in the middle.

Is memorization that bad?

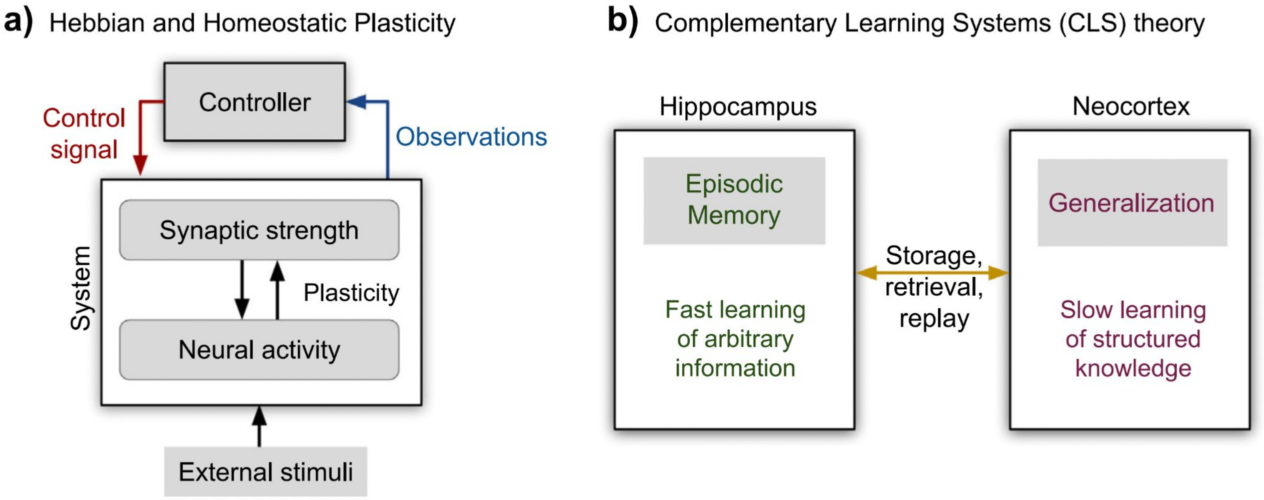
- As at least in various classification tasks outliers are important, we intend to adapt our models training procedure in order to assimilate them properly.
- In fact, we always face the trade-off between abilities to adjust towards new knowledge and the robust generalization of our networks - just as our brains do.


a) Hebbian and Homeostatic Plasticity
Controller
Control signal
Observations
System
Synaptic strength
Plasticity
Neural activity
External stimuli
b) Complementary Learning Systems (CLS) theory
Hippocampus
Episodic Memory
Fast learning of arbitrary information
Storage, retrieval, replay
Neocortex
Generalization
Slow learning of structured knowledge

Parisi et al., 2019

Hebbian and Homeostatic Plasticity Theory

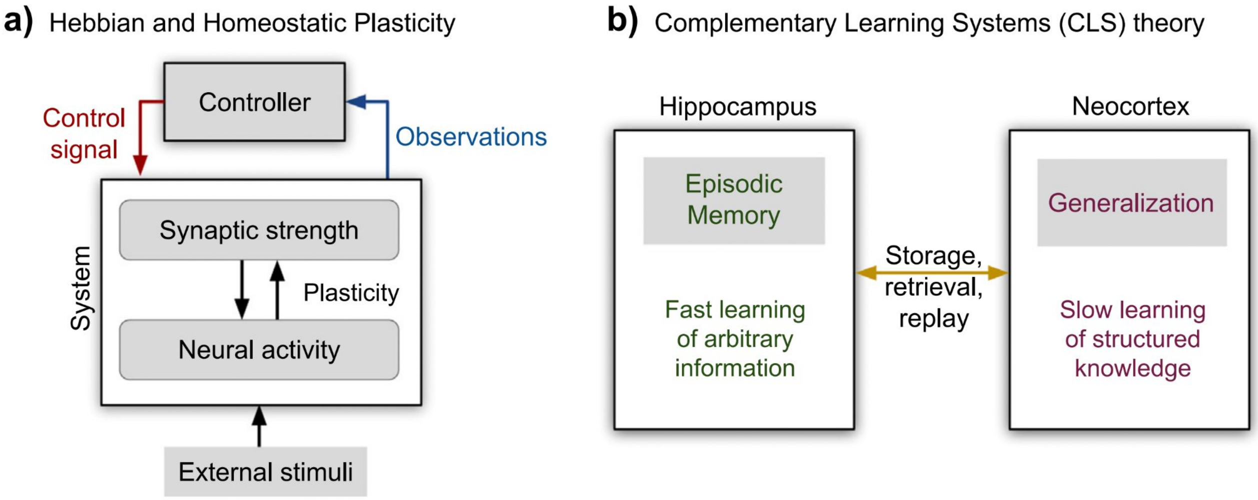
- Hebbian Plasticity postulates the way that our brains strengthen the connection between neurons, if one of them drives the activity of another.
- However, such system would face catastrophic forgetting without Homeostatic Plasticity, which serves as a compensatory feedback controlling the unstable dynamics of the system. For instance, we may consider various regularizations on neuron connections.


a) Hebbian and Homeostatic Plasticity
Controller
Control signal
Observations
System
Synaptic strength
Plasticity
Neural activity
External stimuli
b) Complementary Learning Systems (CLS) theory
Hippocampus
Episodic Memory
Fast learning of arbitrary information
Storage, retrieval, replay
Neocortex
Generalization
Slow learning of structured knowledge

Parisi et al., 2019

Complementary Learning Systems Theory

- Our brains adapt different regions towards rapid memorization of observations in Hippocampus [in sparse representations] and slower generalization among structured memories in Neocortex. Hippocampus does not drop off its influence over time, being responsible for switching between pattern discrimination and completion for recalling information. In other words, our brains do not eliminate the memorization overall.



a) Hebbian and Homeostatic Plasticity

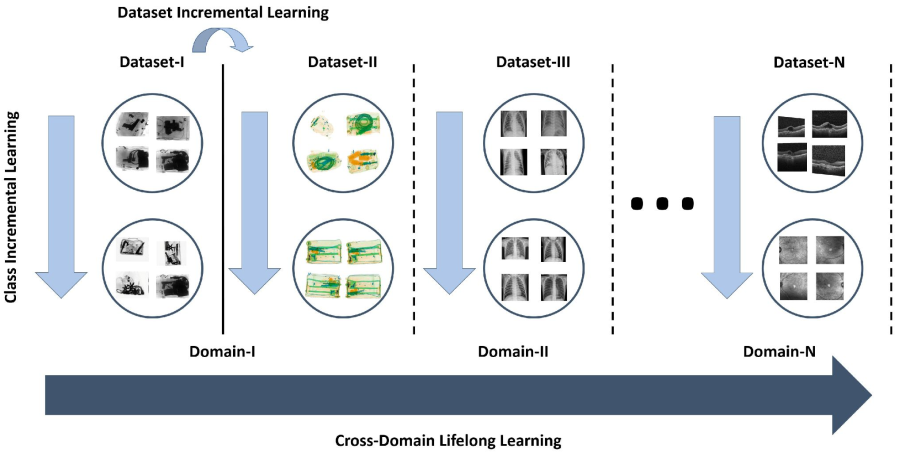
b) Complementary Learning Systems (CLS) theory

Hippocampus
Episodic Memory
Fast learning of arbitrary information
Storage, retrieval, replay
Neocortex
Generalization
Slow learning of structured knowledge

Parisi et al., 2019

Well… we are far from brains

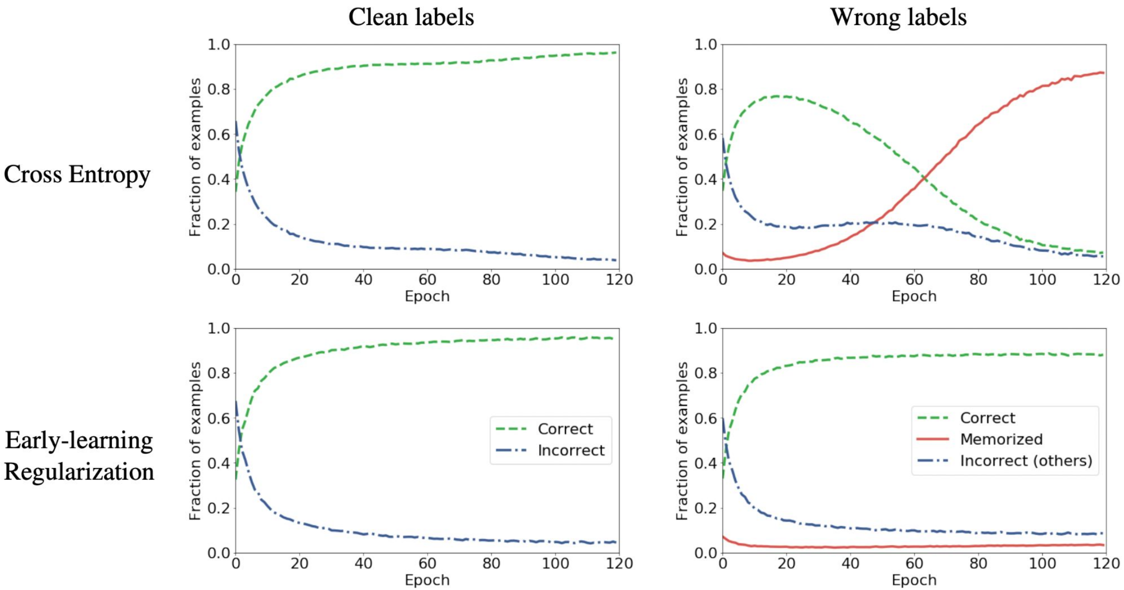
- Neural networks still perform badly in continual lifelong learning tasks, being unstable in terms of acquiring new knowledge. So, in some sense, attempts to bear memorization follow the idea of finding the right balance, as memorization tends to outweigh generalization usually. Let’s now take a closer look at when does it happen.


Dataset Incremental Learning
Dataset-I
Dataset-II
Dataset-III
Dataset-N
Class Incremental Learning
Domain-I
Domain-II
Domain-N
Cross-Domain Lifelong Learning

Khan et al., 2022

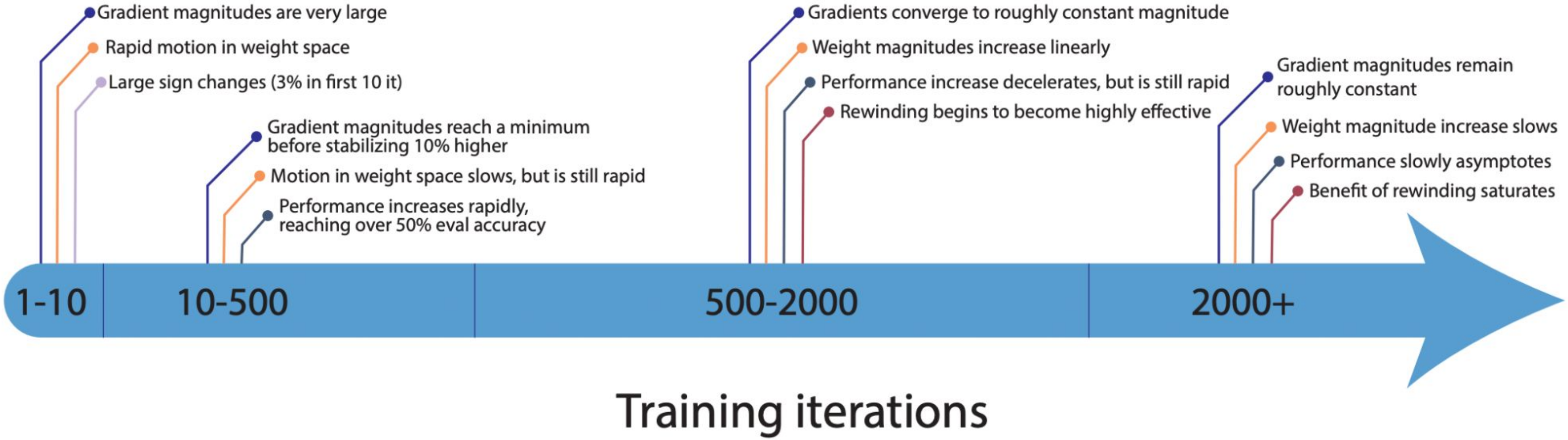
When does the memorization happen?

Looking once more at the process of memorization on wrong labels, we derive that initially models generalize reasonably well. However, they consistently shift towards wrong memorization, which suggests the necessity of regularization during the early stage of learning.



Liu et al., 2020

Early Learning



1-10

- Gradient magnitudes are very large
- Rapid motion in weight space
- Large sign changes (3% in first 10 it)

10-500

- Gradient magnitudes reach a minimum before stabilizing 10% higher
- Motion in weight space slows, but is still rapid
- Performance increases rapidly, reaching over 50% eval accuracy

500-2000

- Gradients converge to roughly constant magnitude
- Weight magnitudes increase linearly
- Performance increase decelerates, but is still rapid
- Rewinding begins to become highly effective

2000+

- Gradient magnitudes remain roughly constant
- Weight magnitude increase slows
- Performance slowly asymptotes
- Benefit of rewinding saturates

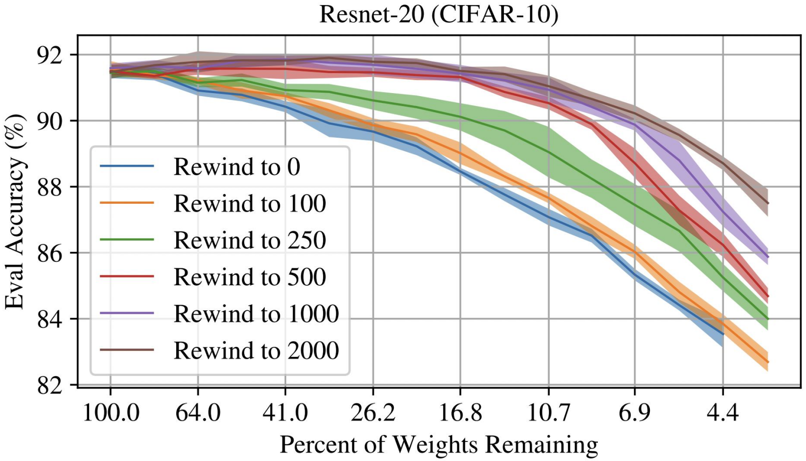
Training iterations

Figure 2: Rough timeline of the early phase of training for ResNet-20 on CIFAR-10.

Frankle et al., 2020

Lottery-Ticket Hypothesis Recap

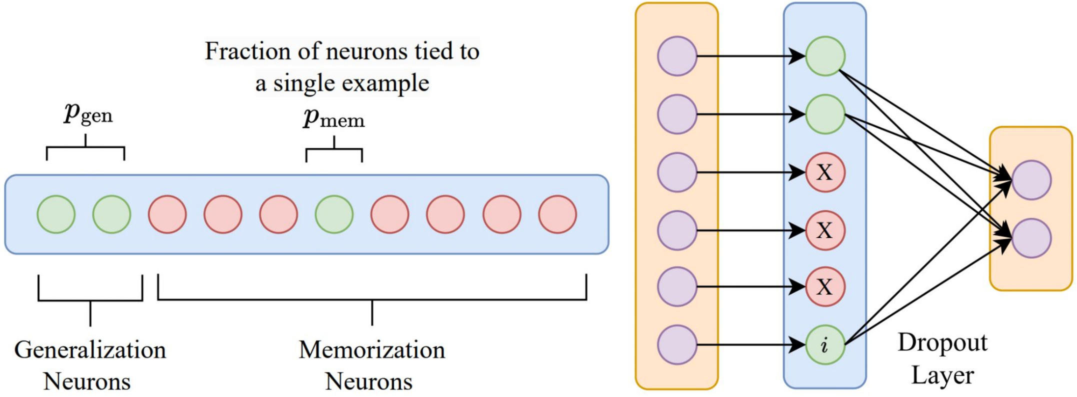
- Train a network on the task without pruning
- Prune top m% of its weights by their magnitude
- Reinitialize them somehow:
    - Rewinding back to iteration 0 - the same random
    - Initializing randomly once more
    - Rewinding back to iteration k
- Train pruned network once more
- Voilà, it performs not worse (or better!)
- *Arbitrary pruned networks wouldn't achieve such results being trained by itself



Frankle et al., 2020

Lottery-Ticket vs Example-tied Dropout ?

- While pruning the lottery-ticket, we select weights with largest magnitudes
- In the example-tied dropout the fraction of memorizing neurons is small per each sample, so they tend to absorb sudden gradient spikes.
- However, as they are being rarely updated, they may not drift away from zero that far, comparing to generalization neurons. Does this mean that we artificially induce generalization neurons to be produced by larger weights?



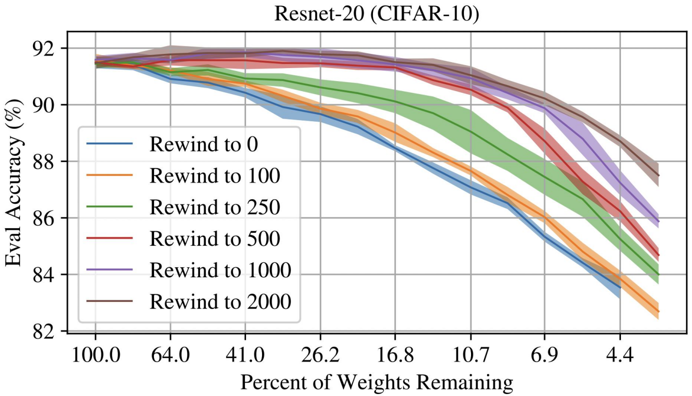
Maini et al., 2023

Lottery-Ticket vs Example-tied Dropout ?

If this behavior is really observed, then example-tied dropout may be performing similar things to model pruning. Therefore, we may look at the success of lottery-tickets rewinding in terms of memorization in such way:

- High magnitude pruning selects more-generalizing weights overall
- Rewinding rolls back generalization of these weights to the better early-learning stage
- Not being tied to memorization, these weights are capable of learning once more

These thoughts may be controversial though*



Frankle et al., 2020

Questions?

Thank you for your attention!

Literature

Maini et al., 2023: Can Neural Network Memorization Be Localized?

Carlini et al., 2019: Distribution Density, Tails, and Outliers in Machine Learning: Metrics and Applications

Parisi et al., 2019: Continual lifelong learning with neural networks: A review

Liu et al., 2020: Early-Learning Regularization Prevents Memorization of Noisy Labels

Frankle et al., 2019: The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks

Frankle et al., 2020: The Early Phase of Neural Network Training