Repeat before Forgetting: Spaced Repetition for Efficient and Effective Training of Neural Networks

Hadi Amiri, Timothy Miller, Guergana Savova
first.last@childrens.harvard.edu
Motivation

• Efficiency challenge
  ▫ Large amount of data is usually required to train effective neural models
  ▫ Long turnaround time
    • E.g. Google’s neural MT system takes three weeks to be trained with 100 GPUs

Wu et al., 2016; Hirschberg and Manning, 2015
Motivation

- Efficiency challenge
  - Large amount of data is usually required to train effective neural models
  - Long turnaround time
    - E.g. Google’s neural MT system takes three weeks to be trained with 100 GPUs

Develop training paradigms that lead to efficient and also effective training of neural models.
Motivation

- **Human Memory**
  - Memory retention / Recall Probability
    - Prob. that a human recalls a previously-seen item declines in time, especially if there is no attempt to retain information

Ebbinghaus, 1913
Motivation

- **Human Memory**
  - Memory retention / Recall Probability

- **Spaced Repetition**
  - Human learners can learn efficiently and effectively by increasing intervals of time between subsequent reviews of previously learned materials.

---

Ebbinghaus, 1913; Dempster, 1989; Cepeda et al., 2006
Motivation

• **Recall Indicators**
  - **Delay** since last review of the item,
    - Lower recall probability for longer delays
  - **Difficulty** of the item,
    - Lower recall probability for more difficult items
  - **Strength** of the human memory
    - Higher recall probability for stronger memories.

Cepeda et al., 2006; Averell and Heathcote, 2011, Reddy et al., 2016
Research Question

• Can we develop memory model that can accurately estimate the time by which an item might be forgotten by a learner?
  ▫ Schedule a review for the learner before that time for effective and efficient learning.
    • More effective: the learner reviews materials before they are forgotten, and
    • More efficient: the learner avoids reviewing materials that he/she already knows.
Research Question

• Can we develop memory model that can accurately estimate the time by which an item might be forgotten by a learner?
  ▫ Schedule a review for the learner before that time for effective and efficient learning.
Key Questions

• **Q1**: Is there an analogy between training neural networks and findings in psychology about human memory model?
Key Questions

• **Q1:** Is there an analogy between training neural networks and findings in psychology about human memory model?

• **Q2:** Could we use spaced repetition to efficiently and effectively train neural networks?
Q1
Is there any analogy between training neural networks and findings in psychology about human memory model?
Q1: Recall Indicators & Net Retention

- Directly evaluate the effect of the recall indicators on memory retention in neural networks

A: base set  80%
B: review set  10%
C: replacement set  10%
Q1: Delay Since Last Review

- Keep recall point fixed
- Move sliding window

**Inverse relation** between network retention and delay since last review in neural networks
Q1: Item Difficulty

- Loss at last review point

**Inverse relation** between network retention and item difficulty
Q1: Network Strength

- Accuracy on validation data at recall point
- Keep delay fixed
- Increase Rec

Direct relation between network retention and network strength
Q1: Recall Indicators & Net Retention

- **Delay Effect**
  - Neural networks forget training examples after a certain period of intervening training data

- **Item Difficulty Effect**
  - The period of recall is shorter for more difficult examples, and

- **Network Strength Effect**
  - Recall improves as networks achieve better overall performance.
Q2
Could we use spaced repetition to efficiently and effectively train neural networks?
Q2: Spaced Repetition

• Challenge
  ▫ Estimating the time by which a training instance should be reviewed!
    • Ideally before it is forgotten,
    • Ideally avoid early reviews.
Q2: Spaced Repetition

• Solution
  ▫ Use recall indicators to lengthen or shorten review intervals with respect to individual learners and training instances
Q2: Spaced Repetition

• Solution
  ▫ Use recall indicators to lengthen or shorten review intervals with respect to individual learners and training instances
    • Delay: epochs to next review of instance
    • Item Difficulty: Loss of network for instance
    • Network Strength: performance on validation data.

\[ x_i = \frac{d_i \times t_i}{s_e} \]

- \( d_i \): loss of network for instance \( i \)
- \( t_i \): epochs to next review of \( i \)
- \( s_e \): performance of network—on validation data—at epoch \( e \).
Q2: Spaced Repetition

- Use density kernels as schedulers that favor less difficult training instances in stronger networks.

\[ x_i = \frac{d_i \times t_i}{s_e} \]

\(d_i\): loss of network for instance \(i\)

\(t_i\): epochs to next review of \(i\)

\(s_e\): performance of network—on validation data—at epoch \(e\).
Q2: Spaced Repetition

Training data
Q2: Spaced Repetition

- **Training data**
- **Current batch**
  - \( t \leq 1 \)
- **Delayed batch**
  - \( t > 1 \)

- **Epoch**
  - \( e \) to \( e+1 \)
Q2: Spaced Repetition

Training data

$\text{Current batch}$

$\hat{t}_i = \arg \min_{t_i} (f(x_i, \hat{t}) - \eta)^2$

$\text{Delayed batch}$

Train network

$\hat{t}_i = t_i - 1$

$e$ $e+1$ epoch
Evaluation

• We evaluate schedulers based on
  ▫ Scheduling accuracy
    • accuracy in estimating network retention with respect to previously-seen instances
  ▫ Effect on the efficiency
    • Time required to train the network (running time)
    • Number of instances used for training per epoch
  ▫ Effect on the effectiveness
    • Accuracy of the downstream network on test data
Scheduling accuracy

Scheduler predicts a delay $t$ for an item, i.e. $t$ epochs to next review of the item.

evaluate network retention with respect to the item at epoch $e + t$

A good scheduler accurately delays more items.
Scheduling accuracy

Rbf kernels accurately delay substantial amount of instances per epoch.
Efficiency & Effectiveness

- Rbf kernels are 2.9-4.8 times faster than standard training.
- Rbf kernels show comparable training accuracy to standard training.
Efficiency & Effectiveness

CL starts small and gradually increases the amount of training data by adding harder instances into its training set.

RbF with high recall confidence thr start big and gradually delays reviewing instances that the networks have learned.
Conclusion

1. Memory retention in neural networks is affected by the same (known) factors that affect memory retention in humans

2. A novel training paradigm for neural networks based on spaced repetition

3. Can be applied without modification to any neural network.
Future Work

- Could training paradigms benefit from each other?
  - Predict easiness of novel training instances to inform CL
  - Incorporate Leitner’s queuing mechanism in CL or RbF.
- Could recall confidence parameter be learned dynamically?
  - E.g. with respect to network behavior?
- Theoretical analysis on their lower and upper bounds
  - More flexible delay functions.
Thank you

• Code
  ▫ [https://scholar.harvard.edu/hadi/RbF](https://scholar.harvard.edu/hadi/RbF)

• Questions
  ▫ [hadi.amiri@childrens.harvard.edu](mailto:hadi.amiri@childrens.harvard.edu)