CMPE489: Cognitive Science - Final Project
Replication Study of "Multitask Learning Via Interleaving: A Neural Network Investigation"

Efekan Kavalcı (efekan.kavalci@std.bogazici.edu.tr)
MS, Department of Computer Engineering - Boğaziçi University

Ramazan Burak Sarıtaş (ramazan.saritas@std.bogazici.edu.tr)
BS, Department of Computer Engineering - Boğaziçi University

Ali Alperen Sönmez (ali.sonmez@std.bogazici.edu.tr)
BS, Department of Computer Engineering - Boğaziçi University

Abstract

This study replicates the findings of "Multitask Learning Via Interleaving: A Neural Network Investigation," focusing on the effects of multitask learning on learning and memory. We use ResNet-18 model as an alternative to ResNet-50, and STL-10 and MNIST datasets as alternatives to CIFAR-10 and SVHN. The training alternates between different tasks, following a specified procedure. We compare our results to the original study, analyzing accuracy metrics, and inspecting task switching effects and memory consolidation. Our findings provide insights into the dynamics of the multitask learning via interleaving concept across various dataset combinations.

Repository of this study can be found at the following GitHub link: github.com/rburaksaritas/cmpe489-final-project

LaTeX of this report can be found at the following Overleaf link: www.overleaf.com/read/znbprfrgwbmg#3464f1

Keywords: multitask learning; neural networks; relearning; switching cost; convolution

Introduction

Background and Motivation

The original study "Multitask Learning Via Interleaving: A Neural Network Investigation" explores the impact of interleaving tasks during neural network training. The study demonstrates that interleaving can improve learning efficiency and memory retention. This concept is significant because it offers potential enhancements in training models on multiple tasks simultaneously, which is increasingly relevant in various applications such as computer vision, autonomous driving, natural language processing, and more.

Key Concepts

To better understand the findings, it is essential to explain two key concepts examined in the study:

Forgetting With Relearning Savings This concept refers to the phenomenon where, after a period of not practicing a task, some forgetting occurs. However, relearning the task is faster and more efficient than the initial learning phase. This is because the underlying memory traces are not completely erased and can be reactivated more easily.

Task Switching Cost Task switching cost is the cognitive or performance cost associated with switching from one task to another. When tasks are interleaved, there is a momentary drop in performance each time a switch occurs, as the system needs to reconfigure itself for the new task. This cost can be quantified in terms of time or errors made immediately after a switch.

Objectives of the Replication Study

The primary objective of this replication study is to verify the findings of the original paper by conducting similar experiments with different datasets and ResNet-18 model. By doing so, we aim to assess the robustness and generalizability of the multitask learning via interleaving approach.

Scope and Focus

Our replication study will focus on reproducing the effects of interleaving training tasks on learning and memory. We will use the STL-10 (as an alternative to CIFAR-10) and MNIST (as an alternative to SVHN) datasets to validate the original findings in a different context. We will compare the performance of interleaved training.

Information about the Original Study

Research Questions

The original study was based on the following research questions:
- How does interleaving tasks during neural network training impact learning efficiency and memory retention?
- Can interleaving improve performance on multiple tasks compared to sequential training?

Participants

The original study did not involve human participants as it focused on machine learning models. The participants, in this context, are the datasets used for training and evaluation.

Design

The design imitated the human learning setting by alternating between tasks frequently, as opposed to the traditional machine learning setting which often completes one task before moving to the next.See Figure 1. To imitate this alternated setting in machine learning, tasks are interleaved with specified run-lengths in the multitask learning enviroment. For example, with run-length 3 and two tasks A and B, A is trained

for 3 epochs, then a task switch occurs to switch training to task B; the process goes on periodically.

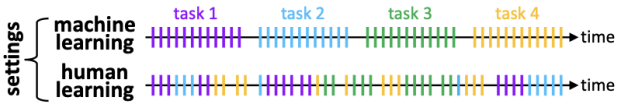

Figure 1: Learning settings of machine and human.

Artifacts

The original study used the following artifacts:
- **Datasets:** CIFAR-10 and SVHN
- **Model:** ResNet-50

Context Variables

Important context variables that affected the design and results include training epochs and batch sizes, learning rates and optimization algorithms, data augmentation techniques.

Summary of Results

The major findings of the original study were:

- Interleaving helped in better memory retention and reduced forgetting.
- The study demonstrated significant task switching costs but also showed relearning savings.

Information about the Replication

Motivation for Conducting the Replication

The primary motivation for conducting this replication study is to validate the results of the original paper "Multitask Learning Via Interleaving: A Neural Network Investigation." Additionally, we aim to broaden the results by using different model and datasets to assess the generalizability and robustness of the findings.

Level of Interaction with the Original Experimenters

This replication is an external replication. The level of interaction with the original researchers was minimal, involving only reading the published paper and accessing publicly available resources such as datasets and model architecture. No direct consultation with the original researchers was conducted.

Changes to the Original Experiment

Several changes were made to the original experiment:

- **Model:** Instead of the ResNet-50 model, we used ResNet-18 due to computational constraints.
- **Datasets:** We replaced CIFAR-10 with STL-10 and SVHN with MNIST to validate the findings in a different context.
- **Training Procedure:** The training procedure was adapted to the ResNet-18 model and the new datasets. Specific details about the optimizer, learning rate, and data augmentation techniques were adjusted to fit the new setup. All experiments are run on a T4 GPU.

The motivation for these changes was primarily to explore the generalizability of the original study's findings with different resources and to address computational limitations.

Replication Experiments

Experiment: CIFAR5 & CIFAR5

This experiment aims to verify whether ResNet-18 can serve as an effective alternative to ResNet-50 for the multitask learning via interleaving methodology that is described in the original paper. The authors use CIFAR-10 dataset in the original study. By using the CIFAR-10 dataset divided into two subsets, each containing 5 classes (CIFAR5 & CIFAR5), we intend to determine whether ResNet-18 performs similar to ResNet-50. This verification is crucial as it will enable us to apply the replication study with different datasets using ResNet-18, which is more computationally feasible.

Experimental Setup

- **Datasets:** CIFAR-10 divided into two subsets (CIFAR5 & CIFAR5).
- **Model:** ResNet-18.
- **Training Procedure:** The model was trained with interleaved tasks using run lengths of 3, 6, and 15 epochs. We use an SGD optimizer with learning rate 0.02, and without momentum, which was the setting in the original paper. Batch size is set as 512 for the training. A total of 90 epochs training is done.
- **Evaluation Metric:** Training accuracy is evaluated at each batch before taking a gradient step for that training batch. Additionally, for each training step a validation batch each for task 1 and task 2 are used to evaluate validation accuracy. Validation batch size is set as 512.

Results See Figure 7 for validation accuracy with run length 3.

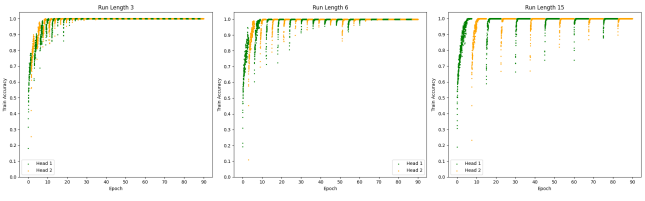

Figure 2: CIFAR5 & CIFAR5 training accuracy under various run lengths.

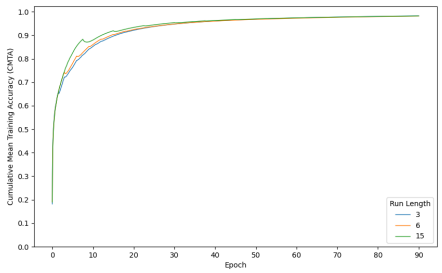

Figure 3: CIFAR5 & CIFAR5 cumulative mean training accuracy (CMTA).

Analysis The results show that the ResNet-18 model performed qualitatively similarly to the ResNet-50 model in terms of training accuracy. The training accuracy vs. epoch plots are given in Figure 2, for run lengths 3, 6, and 15. This qualitative similarity means ResNet-18 can be a good alternative for the experiments that investigate multitask learning via interleaving. It uses less computational power without losing much performance. The phenomena mentioned in the original study such as task switching cost and forgetting with relearning with savings are observed in a similar manner in the replicated results. An important point to note is that ResNet-50 is a relatively larger model than ResNet-18 (25.6 million vs. 11.7 million parameters), thus our replication with ResNet-18 resembles a somewhat faster learning. As a consequence, the training accuracy rises faster and forgetting is less evident during task-switches.

Sincre the replication experiment with ResNet-18 reflects the characteristics with those from the original study using ResNet-50, we apply ResNet-18 in the following experiments where we used different datasets, including STL-10, MNIST, FashionMNIST. We also provide an experiment with CIFAR-100 dataset in the "Further Experiments" section. CIFAR-100 dataset includes 100 classes and is a more challenging dataset in terms of classification, due to the large number of classes.

Experiment: STL5 & STL5

This experiment aims to replicate the multitask learning via interleaving methodology using the STL-10 dataset divided into two subsets, each containing 5 classes (STL5 & STL5). The goal is to assess the impact of interleaving on learning characteristics. STL-10 is a dataset with 5000 training examples and 8000 test examples, the distribution is balanced across the 10 classes. 10 classes in STL-10 are the same with CIFAR-10 dataset. However, STL-10 includes higher resolution images (96x96) compared to CIFAR-10).

Experimental Setup

- **Datasets:** STL-10 divided into two subsets (STL5 & STL5).
- **Model:** ResNet-18.
- **Training Procedure:** Training batch size is set as 64, since STL-10 has significantly lower number of training examples. Remaining settings are the same with the replication experiment (CIFAR5 & CIFAR5).
- **Evaluation Metrics:** Same with the replication experiment (CIFAR5 & CIFAR5). Validation batch-size is set as 256 due to the memory limitations in the GPU.

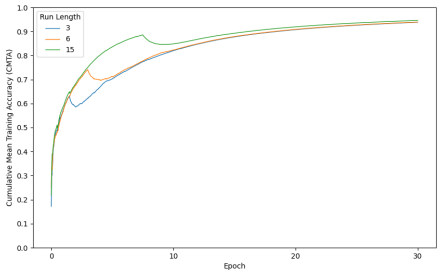

Figure 4: STL5 & STL5 Cumulative Mean Training Accuracy (CMTA).

Results & Analysis The train accuracy results of this experiment has similar general trends with the CIFAR5 & CIFAR5 experiment, as can be seen from Figure 7. Forgetting during task switches and faster recovery due to the relearning savings phenomenon in the later stages can be observed. However, due to the small batch size used in the training, the accuracy values are noisy, we can see unexpectedly low accuracies even in the later stages. The noise is also evident in validation accuracies, however the general trend is same again; the validation accuracy of the training task increases where the other task's validation accuracy decreases in a run. See Figure 8 for validation accuracy with run length 3. The characteristic of cumulative mean training accuracy (CMTA) is similar, less task switches result in higher CMTA.

Experiment: MNIST5 & FASHIONMNIST5

This experiment aims to replicate the multitask learning via interleaving methodology using two different datasets: MNIST and FashionMNIST. The two datasets have 28x28 grayscale images of digits and fashion items, respectively. Thus the nature of two tasks are quite different. This is similar to CIFAR10 and SVHN combination in the original study. However, MNIST and FashionMNIST are considerably easier datasets than CIFAR10 and SVHN, especially for a non-trivial model such as ResNet-18. This allows us to observe how the training behaves when an easy task is applied.

Experimental Setup

- **Datasets:** A subset that includes 5 classes from MNIST and a subset that includes 5 classes from FashionMNIST (MNIST5 & FashionMNIST5).

- **Model:** ResNet-18 model adjusted to grayscale images.
- **Training Procedure:** The settings are the same with the replication experiment (CIFAR5 & CIFAR5).
- **Evaluation Metrics:** Same with the replication experiment (CIFAR5 & CIFAR5).

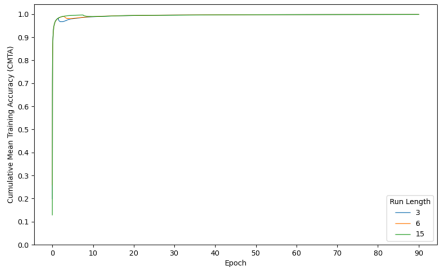

Figure 5: MNIST5 & FashionMNIST5 cumulative mean training Accuracy (CMTA).

Results & Analysis See Figure 7 for training accuracy through run lengths 3, 6, 15 for MNIST5 & FashionMNIST5. Due to the low complexity of tasks, the training accuracy rapidly rises to very high values (about 99%). In the original study, authors report an "extended training" setting where they applied interleaved and blocked training for 600 epochs. They report that both tasks are learned at later stages in the extended training. Although we do not apply such an extended training, results for MNIST5 & FashionMNIST5 resemble the later stages of this extended training setting, both tasks are learned reasonably well, as can be observed from Figure 8 for validation accuracy. The validation accuracy fluctuates as the task switches occur, but this is inevitable to some extent due to the nature of the interleaved training setup.

Further Experiments

Experiment: CIFAR50 & CIFAR50

This experiment aims to test the multitask learning via interleaving methodology with a large number of classes and see how the method generalizes in such a setup.

Experimental Setup

- **Datasets:** CIFAR100 divided into two subsets (CIFAR50 & CIFAR50).
- **Model:** ResNet-18.
- **Training Procedure:** The settings are the same with the replication experiment (CIFAR5 & CIFAR5) except a longer training with 120 epochs is applied.
- **Evaluation Metrics:** Same with the replication experiment (CIFAR5 & CIFAR5) except the validation batch size is set to 1024.

Results & Analysis The training accuracies exhibit very similar characteristics with the original study and CIFAR10 dataset. If we compare with CIFAR5 & CIFAR5 experiment in our study, we observe that forgetting at the task switches are more significant. However, faster recovery trend is also observed in the training. See Figure 7 for training accuracy through run lengths 3, 6, 15. Moreover, validation accuracies are quite low, which is somewhat expected considering the challenging nature of the task. See Figure 8 for validation accuracy with run length 3. The overfitting in this case is significant, and have to be managed to improve the validation

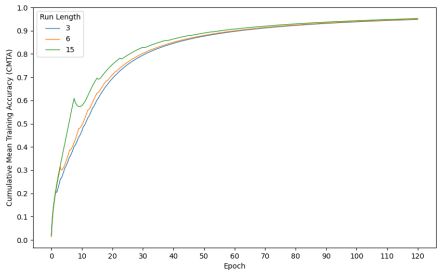

Figure 6: CIFAR50 & CIFAR50 cumulative mean training Accuracy (CMTA).

Comparisons & Conclusions Across Studies

Consistent Results

The replication study produced several results that were consistent with the findings of the original study:

- **Characteristics of Learning with Interleaving:** Similar to the original study, our experiments showed that interleaving tasks during training results in forgetting during task-switches. The forgetting is more evident as run lengths get larger.
- **Relearning Savings:** Both studies observed significant relearning savings when tasks were revisited, indicating that interleaving helps in consolidating knowledge.

Differences in Results

There were some differences between the replication results and the original study. These differences highlight the importance of considering the impact of model architecture and dataset characteristics on multitask learning performance.

- **Task Switching Costs:** Both the original study and our study reports task switching costs, but our experiments showed relatively lower switching costs. This difference could be attributed to the use of different datasets and model. For example, MNIST & FashionMNIST experiment reaches high accuracies rapidly, thus the switching costs are naturally lower. Moreover, utilizing a smaller network such as ResNet-18 allows more faster training due to the smaller number of parameters.

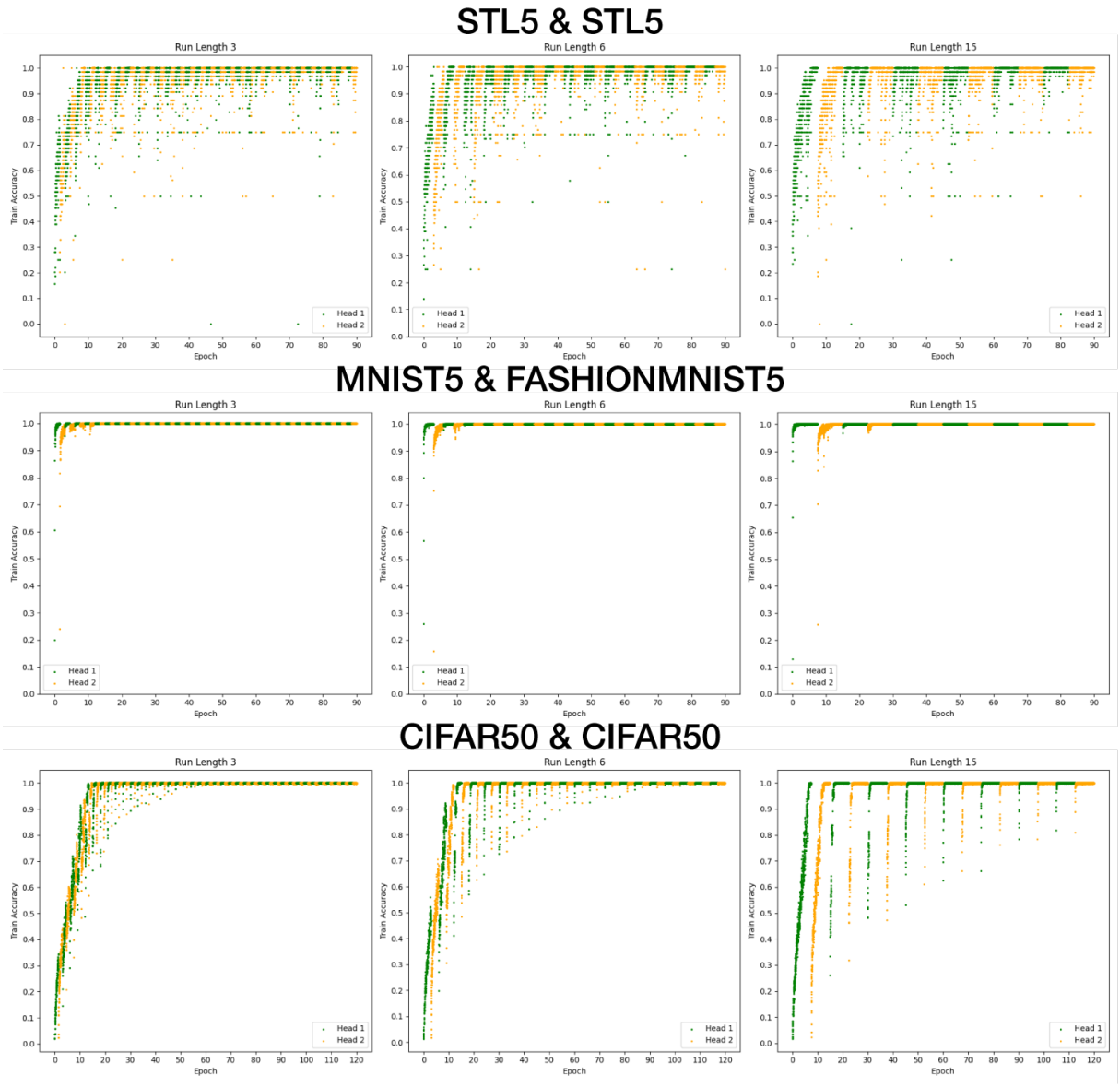

Figure 7: Training Accuracy plots for experiments with run lengths 3, 6 and 15.

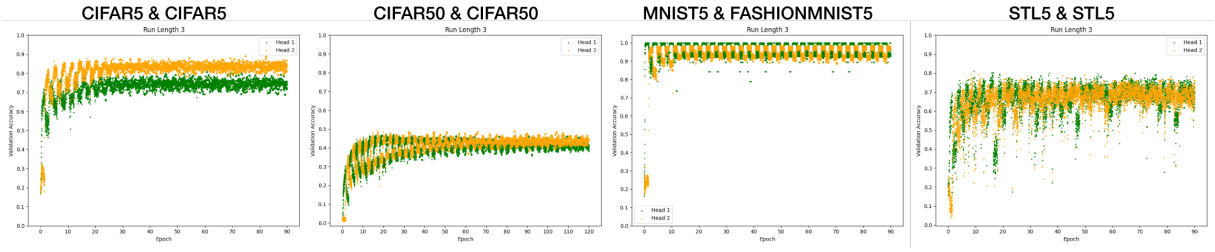

Figure 8: Validation Accuracy plots for experiments with run length 3.

- **Performance Metrics:** The replication showed quantitaviely different results in CIFAR-10 dataset. While the training setting is the same, our model is a smaller model and do not apply several trials of the experiments to account for weight initialization effects, and we do not try lots of combinations in the step of constructing subsets etc. These factors result in quantitatively different results.

Acknowledgments

We would like to thank our instructor Ayşe Başar for their guidance and support throughout this project. We also appreciate the original researchers of "Multitask Learning Via Interleaving: A Neural Network Investigation" - David Mayo, Tyler R. Scott, Mengye Ren, Gamaledin Elsayed, Katherine Hermann, Matt Jones, and Michael Mozer - for their work that inspired this replication study. Additionally, we acknowledge ChatGPT for assisting in generating plots used in this report.

References

Coates, A., Ng, A., & Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. *AISTATS*. Retrieved from https://cs.stanford.edu/~acoates/stl10/

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*. Retrieved from https://arxiv.org/abs/1512.03385

Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images.* Retrieved from https://www.cs.toronto.edu/~kriz/cifar.html

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. Retrieved from http://yann.lecun.com/exdb/mnist/

Mayo, D., Scott, T. R., Ren, M., Elsayed, G., Hermann, K., Jones, M., & Mozer, M. (2023). *Multitask learning via interleaving: A neural network investigation* (Technical Report). University of California. Retrieved from https://escholarship.org/uc/item/3tb956hb

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., & Ng, A. Y. (2011). *Reading digits in natural images with unsupervised feature learning.* Retrieved from http://ufldl.stanford.edu/housenumbers/