Faster Neural Network Training with Data Echoing

Dami Choi, Alexandre Passos, Christopher J. Shallue, George E. Dahl
Google Brain
{damichoi,apassos,shallue,gdahl}@google.com

Abstract

In the twilight of Moore's law, GPUs and other specialized hardware accelerators have dramatically sped up neural network training. However, earlier stages of the training pipeline, such as disk I/O and data preprocessing, do not run on accelerators. As accelerators continue to improve, these earlier stages will increasingly become the bottleneck. In this paper, we introduce "data echoing," which reduces the total computation used by earlier pipeline stages and speeds up training whenever computation upstream from accelerators dominates the training time. Data echoing reuses (or "echoes") intermediate outputs from earlier pipeline stages in order to reclaim idle capacity. We investigate the behavior of different data echoing algorithms on various workloads, for various amounts of echoing, and for various batch sizes. We find that in all settings, at least one data echoing algorithm can match the baseline's predictive performance using less upstream computation. In some cases, data echoing can even compensate for a 4x slower input pipeline.

1 Introduction

Over the past decade, dramatic increases in neural network training speed have facilitated dramatic improvements in predictive performance by allowing researchers to train bigger models using larger datasets and to explore new ideas more rapidly. As Moore's law ends, general purpose processors are no longer rapidly becoming faster, but specialized hardware continues to drive significant speedups by optimizing for a narrower set of operations. For example, GPUs and TPUs[1] optimize for highly parallelizable matrix operations, which are core components of neural network training algorithms.

However, training a neural network requires more than just the operations that run well on accelerators, so we cannot rely on accelerator improvements alone to keep producing speedups in all cases. A training program may need to read and decompress training data, shuffle it, batch it, and even transform or augment it. These steps may exercise multiple system components, including CPUs, disks, network bandwidth, and memory bandwidth. It is impractical to design specialized hardware for all these general operations that involve so many different components. Meanwhile, accelerator improvements are outpacing improvements in general purpose computation, and there already exist workloads where the code running on accelerators consumes only a small portion of the overall wall time. Therefore, if we want to continue to make neural network training faster, we must either (1) make the non-accelerator work faster, or (2) reduce the amount of non-accelerator work required to achieve the desired predictive performance. Although option (1) is appealing, it might require substantial engineering labor or techniques that do not generalize easily (e.g. Zhu et al., 2018; Ying et al., 2018; Kumar et al., 2018). Even if possible, adding more workers might be too expensive or add too much complexity to the system. Instead, we focus on option (2) and explore techniques for reducing the total amount of work performed in earlier stages of the training pipeline.

We can view a neural network training program as a data pipeline that buffers and overlaps computations. For example, Figure 1 shows a typical training pipeline for minibatch stochastic gradient

[1] https://www.blog.google/products/google-cloud/google-cloud-offer-tpus-machine-learning/

Preprint. Under review.

Figure 1: A typical neural network training pipeline.

(a) Without data echoing, downstream computational capacity is idle 50% of the time.

(b) Data echoing with echoing factor 2 reclaims downstream computational capacity.

Figure 2: The overlapping computation time for pipeline stages upstream and downstream of the data echoing insertion point, if stages are executed in parallel and $t_{\text{upstream}} = 2t_{\text{downstream}}$.

descent (SGD) and its variants, which are the *de facto* standard algorithms for training neural networks. The training program first reads and decodes the input data, then shuffles the data, applies a set of transformations to augment the data, and gathers examples into batches. Finally, the program iteratively updates the neural network's parameters to reduce its loss function over successive batches; we call this stage the "SGD update" regardless of which SGD variant is used. Alternative orderings of these stages are also possible, and some pipelines might omit stages or add new ones. Since the output of any pipeline stage can be buffered, the computations in different stages overlap and the slowest stage dominates training time. Modern accelerators accentuate any slowness of other pipeline stages by making the optimized matrix operations less likely to dominate the total training time.

In this work, we study ways to speed up neural network training by reducing the total time spent in the earlier part of the training pipeline when this time exceeds the time spent in the latter part of the pipeline (see Figure 2a). This scenario is realistic today and will become more common if accelerator improvements continue to outpace general purpose CPUs and I/O. Specifically, we consider algorithms that reuse the outputs of the first parts of the pipeline for multiple SGD updates to utilize idle processing capacity. We call this general class of algorithms **data echoing** and the number of times each intermediate output gets used the **echoing factor**.

A data echoing algorithm inserts a repeat stage (optionally shuffling) somewhere in the training pipeline before the SGD update. Provided the time taken by upstream tasks (before the repeat stage) exceeds the time taken by downstream tasks (after the repeat stage), this technique will reclaim idle downstream compute capacity and increase the rate of SGD updates to the model (see Figure 2b). Different data echoing algorithms with different behaviors can be implemented by varying the choice of insertion point in the pipeline, the echoing factor, and the amount of shuffling.

In this paper, we demonstrate that:

1. data echoing reduces the amount of upstream computation needed to reach a competitive out-of-sample error rate on various datasets and model architectures;
2. data echoing can support a wide range of echoing factors;
3. the effectiveness of data echoing depends on the insertion point in the training pipeline;
4. data echoing can benefit from additional shuffling after echoing, but does not require it; and
5. countering expectations, data echoing reaches the same error rate as well-tuned baselines.

1.1 Related work

Data echoing shares similarities with experience replay (Mnih et al., 2015), which samples batches from a buffer containing a reinforcement learning agent's past experiences to prevent the most recent interactions from dominating the updates. Although both data echoing and experience replay reuse previous data, our implementation of data echoing chooses the number of times to repeat each example, whereas most implementations of experience replay do not control this explicitly. In addition, local SGD algorithms (Zinkevich et al., 2010; Zhang et al., 2016), which perform multiple local model updates before communicating globally, can be viewed as reusing data to save computation. However, local SGD targets communication overhead between workers and thus is orthogonal to data echoing.

We are aware of two previous papers that describe variants of data echoing. Fischetti et al. (2018) describe a special case of data echoing they call "minibatch persistency" that reuses minibatches for multiple consecutive SGD updates. They run experiments on CIFAR-10, but do not tune metaparameters for the baseline or for their method. Neither their method nor their baseline reach competitive test set numbers in their experiments, leaving open the question of whether minibatch persistency has an advantage over a well-tuned baseline. Similarly, Hoffer et al. (2019) describe a special case of data echoing they call "batch augmentation" that repeats examples multiple times within a given batch, but with different augmentations. None of their experiments tune optimization metaparameters, although their baselines use settings taken from the original papers that introduced each model. Both Fischetti et al. (2018) and Hoffer et al. (2019) primarily motivate their work as methods to improve generalization, only tangentially mentioning the possibility of reclaiming idle computational capacity. We would not expect data echoing to improve generalization for a fixed number of SGD updates, since then repeated data would be *more* valuable than fresh data. Our experiments in Section 3 only show that data echoing can achieve better out-of-sample error for the same amount of *fresh* data.

2 Data Echoing

We implement data echoing by inserting a stage in the training pipeline that repeats (echoes) the outputs of the previous stage. In some cases, we also shuffle the outputs of the echoing stage, but this can require additional memory. If the overhead of repeating data is negligible and the stages on either side of echoing are executed in parallel (e.g. Chien et al., 2018), then the average time for data echoing to complete one upstream step and e downstream steps is

$$\max\left\{t_{\text{upstream}}, e \times t_{\text{downstream}}\right\}, \tag{1}$$

where t_{upstream} is the time taken by all stages upstream of echoing, $t_{\text{downstream}}$ is the time taken by all stages downstream of echoing, and e is the echoing factor. We assume that $t_{\text{upstream}} \geq t_{\text{downstream}}$, since this is the primary motivation for using data echoing. If we denote the ratio of upstream-to-downstream processing time by $R = t_{\text{upstream}}/t_{\text{downstream}}$, then the time to complete one upstream step and e downstream steps is the same for all echoing factors e up to R. In other words, the additional downstream steps are "free" because they utilize idle downstream capacity. Non-integral echoing factors can be achieved in expectation by probabilistically repeating data items.

When using data echoing, the reduction in training time (if any) depends on the trade-off between upstream steps and downstream steps. On the one hand, since repeated data might be less valuable than completely fresh data, data echoing might require more downstream SGD updates to reach the desired predictive performance. On the other hand, data echoing requires only $1/e$ (instead of 1) upstream steps per downstream step. The total number of upstream steps (and thus the total training time) will decrease if the required number of downstream steps increases by a factor less than e. R is the maximum plausible speedup from data echoing, which would be achieved if $e = R$ and repeated data were just as valuable as fresh data.

Given that every operation in the training pipeline takes some time to execute, R is largest if data echoing is applied just before the SGD update, which would result in the same batch being used multiple times per epoch. However, despite the overhead, we might prefer to insert data echoing earlier in the pipeline if it provides a more favorable trade-off between the number of upstream steps and downstream steps. In particular, the following factors influence the behavior of data echoing at different insertion points:

Echoing before or after batching: Echoing before batching means data is repeated and shuffled at the example level instead of the batch level. This increases the likelihood that nearby batches will be different, at the expense of potentially duplicating examples within a batch. Whether diversification across batches or within batches is more important is an empirical question that we address in Section 3. We call the class of algorithms that echo before batching *example echoing* and the class of algorithms that echo after batching *batch echoing.*

Echoing before or after augmentation: Echoing before data augmentation allows repeated data to be transformed differently, potentially making repeated data more akin to fresh data. Methods like dropout that add noise during the SGD update can similarly make repeated data appear different (Hoffer et al., 2019), even in the absence of augmentation or when echoing after augmentation.

The behavior of data echoing is also influenced by the amount of shuffling (if any) performed after the echoing stage. Where applicable, we implement the shuffling stage as a buffer from which items

Table 1: Tasks summary.

Model	Dataset(s)	Task	Evaluation metric	Target
Transformer	LM1B, Common Crawl	Language modeling	Cross entropy	3.9
ResNet-32	CIFAR-10	Image classification	Accuracy	91%
ResNet-50	ImageNet	Image classification	Accuracy	75%
SSD	COCO	Object detection	mAP	0.24

are randomly sampled by the proceeding pipeline stage. The larger the buffer size, the more repeated data are shuffled, and the closer the training algorithm approximates a program that loads the entire training set in memory before sampling data at random. A relatively large buffer size might be quite realistic for some workloads, but we are primarily interested in the case where we can only afford a buffer size that is a relatively small fraction of the (augmented) dataset size.

3 Experiments

We evaluated data echoing on two language modeling tasks, two image classification tasks, and one object detection task. For language modeling, we trained the Transformer model (Vaswani et al., 2017) on the LM1B (Chelba et al., 2014) and Common Crawl[2] datasets. For image classification, we trained ResNet-32 (He et al., 2016) on the CIFAR-10 dataset (Krizhevsky and Hinton, 2009), and ResNet-50 on the ImageNet dataset (Russakovsky et al., 2015). For object detection, we trained the Single Shot Detector (SSD, Liu et al., 2016) on the COCO dataset (Lin et al., 2014).

The primary question we investigated was whether data echoing could provide a training speedup. We measured training time as the number of "fresh" training examples[3] required to reach a target out-of-sample metric value. The number of fresh examples is proportional to the number of upstream steps in the training pipeline, and therefore proportional to wall time if the echoing factor is less than or equal to R, the ratio of upstream-to-downstream processing time (see Section 2). We did not assume or measure the value of R in our experiments, since R depends on the implementation and is likely to change with future hardware developments. Not all of our tasks satisfied $R \geq 1$ in our implementation. Instead, we designed our experiments to investigate whether data echoing could reduce the number of fresh examples needed across various tasks, since this measurement is implementation independent.

For each workload, we ran an initial set of experiments without data echoing and tuned the metaparameters to achieve the best out-of-sample performance within a practical computational budget.[4] We selected the target metric value to be slightly worse than the best observed in the initial experiments to ensure it could be reached reliably. We verified that small changes to our targets did not affect our conclusions. Table 1 summarizes the workloads and target metric values we used in our experiments.

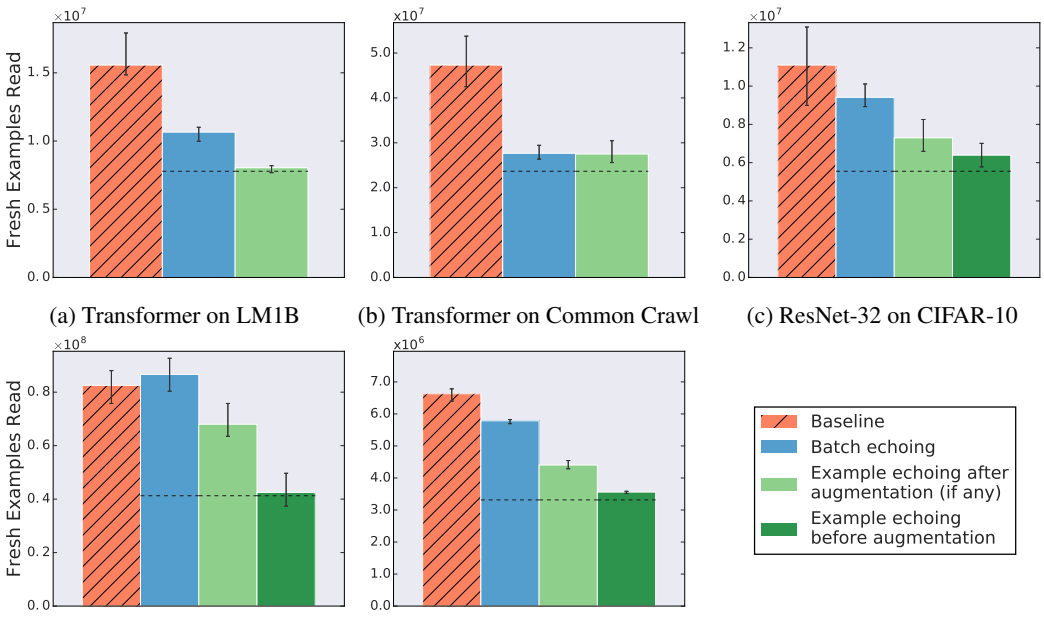
We trained the SSD model using SGD with momentum (Polyak, 1964; Rumelhart et al., 1986) and the Transformer and ResNet models using Nesterov momentum (Nesterov, 1983; Sutskever et al., 2013). We used a constant learning rate for Transformer, and we used learning rate schedules for ResNet (linear decay) and SSD (linear warmup followed by piecewise exponential decay). We preprocessed the text datasets identically to Shallue et al. (2018). We augmented the image datasets at training time by resizing each image, taking a random crop, and randomly horizontally reflecting the cropped images. We randomly distorted the image colors for ImageNet and COCO. Unless otherwise specified, we used a batch size of 1024 for Transformer and ResNet-50, 128 for ResNet-32, and 256 for SSD. We used batch normalization (Ioffe and Szegedy, 2015) for ResNet-50 and SSD with virtual batch sizes (Hoffer et al., 2017) of 32 and 128, respectively.

In each experiment, we independently tuned the learning rate, momentum, and, where applicable, the parameters governing the learning rate schedule. We manually chose the search spaces based on our

[2] `http://commoncrawl.org/2017/07/june-2017-crawl-archive-now-available/`

[3] Each time a training example is read from disk, it counts as a fresh example.

[4] 20k steps for LM1B, 60k for Common Crawl, 110k for ImageNet, 150k for CIFAR-10, and 30k for COCO.



(a) Transformer on LM1B (b) Transformer on Common Crawl (c) ResNet-32 on CIFAR-10

(d) ResNet-50 on ImageNet (e) SSD on COCO

Figure 3: Data echoing with echoing factor 2 either reduces or does not change the number of fresh examples needed to reach the target out-of-sample performance. Dashed lines indicate the expected values if repeated examples were as useful as fresh examples.

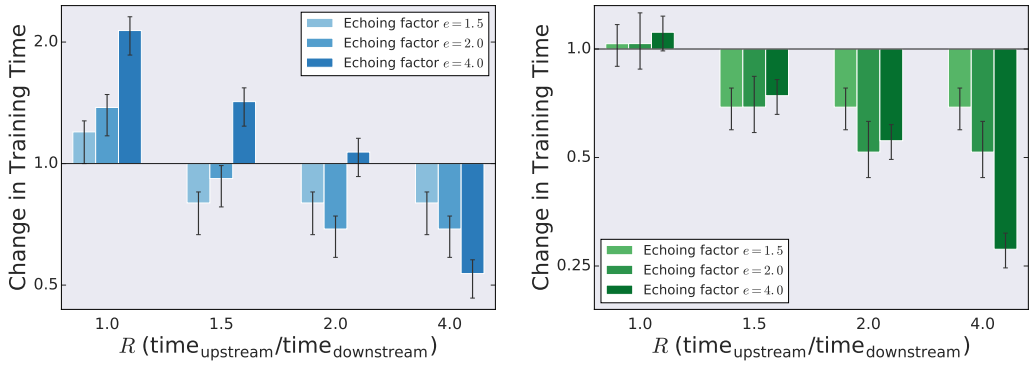
initial experiments, and we verified after each experiment that the optimal metaparameter values were away from the search space boundaries. We used quasi-random search (Bousquet et al., 2017) to tune the metaparameters with fixed budgets of non-divergent[5] trials (100 for Transformer and ResNet-32, and 50 for the more expensive ResNet-50 and SSD models). We then chose the trial that reached the target metric value using the fewest number of fresh examples. We repeated this metaparameter search 5 times for each search space. All figures in this section show the mean number of fresh examples required over these 5 experiments, with the minimum and maximum shown as error bars.

Our experiments evaluated the effects of adding data echoing to a typical neural network training pipeline (Figure 1). We considered three variants of data echoing: example echoing before augmentation, example echoing after augmentation, and batch echoing. For the example echoing variants, we omitted the baseline's "shuffle examples" buffer and inserted a shuffle buffer after the echoing stage with the same size as the baseline's buffer. For batch echoing, we kept the baseline's shuffle buffer and repeated batches without shuffling after the "batch examples" stage. Therefore, our training pipeline always had one shuffle buffer with the same size in all cases, so all data echoing variants used the same amount of memory as the baseline. We used buffer sizes of 10^6 for LM1B and Common Crawl, 10^4 for CIFAR-10, 10^5 for ImageNet, and 10^4 for COCO. We explored the effects of increasing the buffer sizes in Section 3.5.

3.1 Data echoing can reduce the number of fresh examples required for training

Figure 3 shows the effect of data echoing with echoing factor 2 for all workloads in Table 1. In all but one case, data echoing requires strictly fewer fresh examples than the baseline to reach the target out-of-sample performance. The sole exception (batch echoing on ResNet-50) requires about the same number of fresh examples as the baseline – data echoing provides no benefit, but does not harm training either. The earlier echoing is inserted in the pipeline, the fewer fresh examples are needed: example echoing requires fewer fresh examples than batch echoing, and echoing before data augmentation requires fewer fresh examples than echoing after. We did not observe any negative interaction between data echoing and batch normalization for ResNet-50 or SSD.

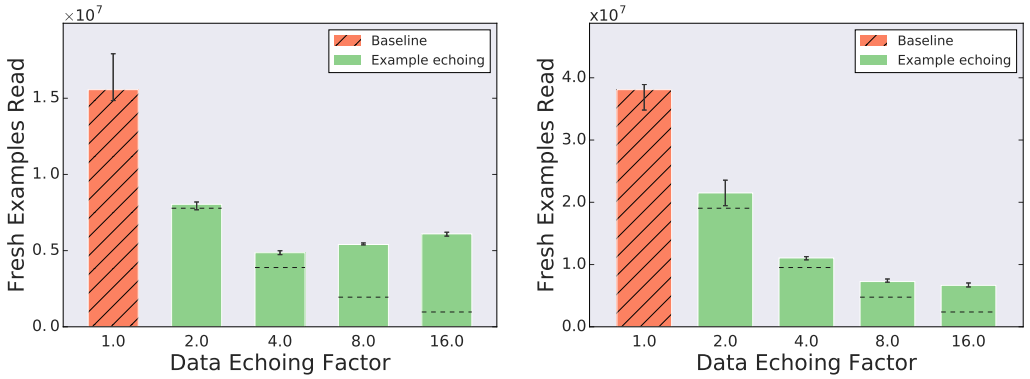
[5]We discarded trials with a divergent training loss, which occurred when the learning rate was too high.



(a) Transformer on LM1B with batch echoing

(b) ResNet-50 on ImageNet with example echoing (before augmentation)

Figure 4: The fractional change in training time (compared to the baseline) for different values of R, the ratio of upstream-to-downstream computation time. Lower is better. The y-axis is in log scale.



(a) Batch size 1024

(b) Batch size 4096

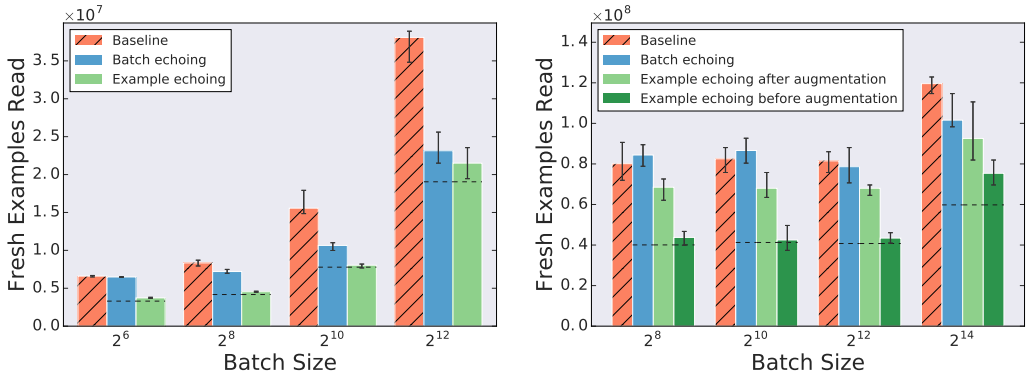
Figure 5: Example echoing reduces the number of fresh examples needed for Transformer on LM1B for echoing factors up to (at least) 16. Dashed lines indicate the expected values if repeated examples were as useful as fresh examples.

3.2 Data echoing can reduce training time

Figure 4 shows the fractional change in training time (assuming negligible overhead from echoing) with various echoing factors e for different values of R, the ratio of upstream-to-downstream processing time. If $R = 1$, data echoing either increases or does not significantly change the training time, as expected. If $R > 1$, any choice of $e \leq R$ reduces training time with the greatest speedup achieved by setting $e = R$. Setting $e > R$ does not reduce training time for Transformer on LM1B, but it does provide a speedup for ResNet-50 on ImageNet, even when $e = 4$ and $R = 1.5$. These results show that data echoing can reduce training time for a range of echoing factors around the optimum value of $e = R$, especially for $e \leq R$.

3.3 Data echoing can be useful up to a reasonable upper bound on the echoing factor

Figure 5 shows the effect of example echoing with echoing factors up to 16 for Transformer on LM1B. For batch size 1024, the maximum useful echoing factor is somewhere between 4 and 8; beyond this value, the number of fresh examples required is larger than for smaller echoing factors. As the echoing factor increases, the number of fresh examples required must eventually exceed the baseline, but even an echoing factor as large as 16 still requires significantly fewer fresh examples than the baseline. For batch size 4096, the maximum useful echoing factor is even larger than 16, suggesting that larger batch sizes can support larger echoing factors than smaller batch sizes.



(a) Transformer on LM1B

(b) ResNet-50 on ImageNet

Figure 6: As the batch size increases, the performance of batch echoing relative to the baseline either stays the same or improves, while for example echoing it either stays the same or gets worse. Dashed lines indicate the expected values if repeated examples were as useful as fresh examples.

3.4 Data echoing as batch size increases

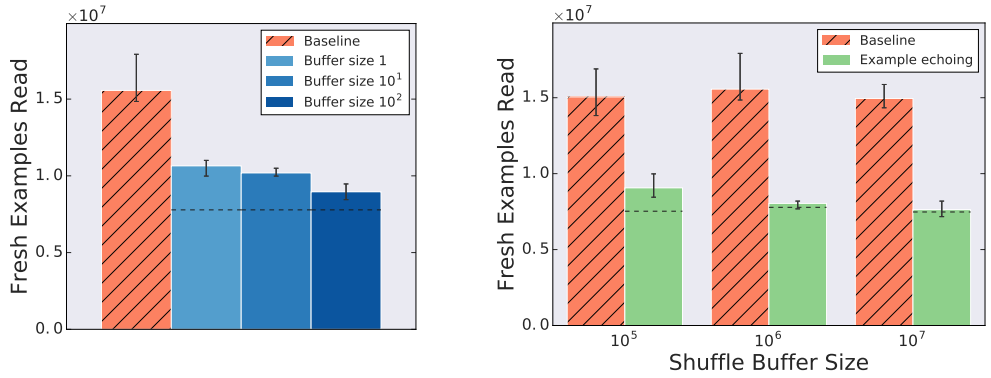
With larger batch sizes, batch echoing performs better, but example echoing sometimes requires more shuffling. Figure 6 shows the effect of data echoing with echoing factor 2 for different batch sizes. As the batch size increases, the performance of batch echoing relative to the baseline either stays the same or improves. This effect makes sense given that repeated batches should approximate fresh batches as the batch size approaches the training set size, and so, in the limit, batch echoing must reduce the required number of fresh examples by the echoing factor. On the other hand, Figure 6 shows that the performance of example echoing relative to the baseline either stays the same or gets worse as the batch size increases. Since the expected fraction of duplicate examples within each batch increases with the batch size, example echoing with larger batches may behave more like a smaller batch size in practice. A smaller batch size may increase the required number of SGD updates (Shallue et al., 2018), which could explain the example echoing results in Figure 6. Increasing the amount of shuffling for repeated examples (at the cost of additional memory) could improve the performance of example echoing at larger batch sizes by reducing the probability of duplicate examples in each batch.

3.5 Data echoing performs better with more shuffling

Figure 7 shows the effect of increasing the shuffle buffer size (at the cost of additional memory) for data echoing with echoing factor 2. While all batch echoing experiments in the previous sections repeated batches without shuffling, the performance of batch echoing improves if repeated batches are shuffled, with more shuffling giving increasingly better performance. Similarly, the performance of example echoing improves with increasing shuffle buffer size, even though it does not help the baseline. This is because more shuffling reduces the probability of duplicate examples within each batch, as discussed in Section 3.4.

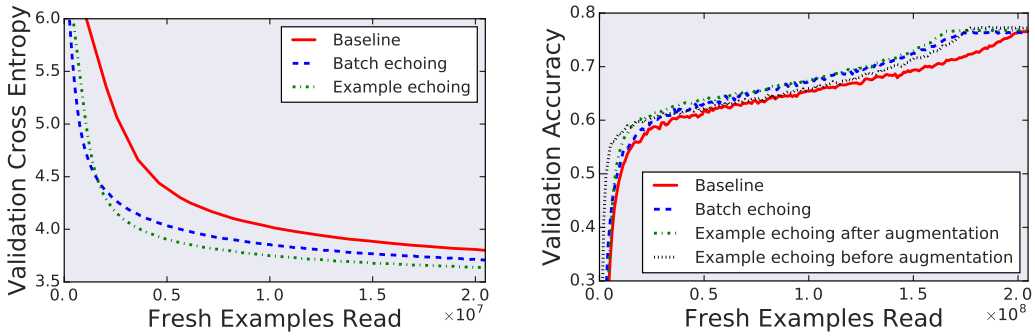
3.6 Data echoing does not harm predictive performance

Although one might be concerned that reusing data could harm final predictive performance, we did not observe any case where data echoing with a reasonable echoing factor failed to reach our target metric value. To further demonstrate that data echoing does not degrade solution quality, we ran experiments with Transformer on LM1B and ResNet-50 on ImageNet to find the best achievable performance within a fixed budget of fresh examples, both with and without data echoing. We picked the fresh-examples budgets so that the baseline models would achieve at least our target metric values from Table 1. We used an echoing factor of 4 for all data echoing experiments. We tuned the metaparameters for the baseline and for all data echoing variants using 500 trials for Transformer and 100 trials for ResNet-50. Figure 8 shows the trials that reached the best out-of-sample performance at



(a) Transformer on LM1B with batch echoing (b) Transformer on LM1B with example echoing

Figure 7: Data echoing performs better with more shuffling. Dashed lines indicate the expected values if repeated examples were as useful as fresh examples.



(a) Transformer on LM1B (b) ResNet-50 on ImageNet

Figure 8: Individual trials that achieved the best out-of-sample performance during training.

any point during training for each experiment. All data echoing variants achieved at least the same performance as the baseline for both tasks.

4 Conclusion

Data echoing is a simple strategy for increasing hardware utilization when the training pipeline has a bottleneck in one of the upstream stages. Although *a priori* one might worry that SGD updates with repeated data would be useless or even harmful, for every workload we considered, at least one variant of data echoing reduced the total number of examples we needed to read from disk. This was true even for Transformer on Common Crawl, a dataset so large that we do not even train for a full epoch. In this case, data echoing reached the target predictive performance while seeing only a subset of the examples seen by the baseline. Echoing after augmentation was still effective at reducing the total number of examples read from disk, making it appealing for image datasets that employ expensive data augmentation that runs on the CPU.

Data echoing is an effective alternative to optimizing the training pipeline or adding additional workers to perform upstream data processing, which may not always be possible or desirable. Although the exact speedup depends on the model architecture, dataset, batch size, and how well repeated data are shuffled, setting the echoing factor to the ratio of upstream-to-downstream processing time maximizes the potential speedup and worked well in our experiments, even for large ratios. As improvements in specialized accelerators like GPUs and TPUs continue to outpace general purpose computation, we expect data echoing and similar strategies to become increasingly important parts of the neural network training toolkit.

References

Olivier Bousquet, Sylvain Gelly, Karol Kurach, Olivier Teytaud, and Damien Vincent. Critical hyper-parameters: No random, no cry. *arXiv preprint arXiv:1706.03200*, 2017.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. In *Conference of the International Speech Communication Association*, 2014.

Steven WD Chien, Stefano Markidis, Chaitanya Prasad Sishtla, Luis Santos, Pawel Herman, Sai Narasimhamurthy, and Erwin Laure. Characterizing deep-learning I/O workloads in TensorFlow. *arXiv preprint arXiv:1810.03035*, 2018.

Matteo Fischetti, Iacopo Mandatelli, and Domenico Salvagnin. Faster SGD training by minibatch persistency. *arXiv preprint arXiv:1806.07353*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pages 1731–1741, 2017.

Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: better training with larger batches. *arXiv preprint arXiv:1901.09335*, 2019.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Sameer Kumar, Dheeraj Sreedhar, Vaibhav Saxena, Yogish Sabharwal, and Ashish Verma. Efficient training of convolutional neural nets on large distributed systems. *2018 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 392–401, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533, 1986.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Christopher J Shallue, Jaehoon Lee, Joe Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E Dahl. Measuring the effects of data parallelism on neural network training. *arXiv preprint arXiv:1811.03600*, 2018.

Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Chris Ying, Sameer Kumar, Dehao Chen, Tao Wang, and Youlong Cheng. Image classification at supercomputer scale. *arXiv preprint arXiv:1811.06992*, 2018.

Jian Zhang, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Parallel SGD: When does averaging help? In *International Conference on Machine Learning Workshop on Optimization in Machine Learning*, 2016.

Yue Zhu, Fahim Chowdhury, Huansong Fu, Adam Moody, Kathryn Mohror, Kento Sato, and Weikuan Yu. Entropy-aware I/O pipelining for large-scale deep learning on HPC systems. In *2018 IEEE 26th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 145–156. IEEE, 2018.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.