
Training Kinetics in 15 Minutes: Large-scale Distributed Training on Videos

Ji Lin
MIT
jilin@mit.edu

Chuang Gan
MIT-IBM Watson AI Lab
ganchuang@csail.mit.edu

Song Han
MIT
songhan@mit.edu

Abstract

Deep video recognition is more computationally expensive than image recognition, especially on large-scale datasets like Kinetics [1]. Therefore, training scalability is essential to handle a large amount of videos. In this paper, we study the factors that impact the training scalability of video networks. We recognize three bottlenecks, including data loading (data movement from disk to GPU), communication (data movement over networking), and computation FLOPs. We propose three design guidelines to improve the scalability: (1) fewer FLOPs and hardware-friendly operator to increase the *computation efficiency*; (2) fewer input frames to reduce the data movement and increase the *data loading efficiency*; (3) smaller model size to reduce the networking traffic and increase the *networking efficiency*. With these guidelines, we designed a new operator Temporal Shift Module (TSM) that is efficient and scalable for distributed training. TSM model can achieve $1.8\times$ higher throughput compared to previous I3D models. We scale up the training of the TSM model to 1,536 GPUs, with a mini-batch of 12,288 video clips/98,304 images, without losing the accuracy. With such hardware-aware model design, we are able to scale up the training on Summit supercomputer and reduce the training time on Kinetics dataset from 49 hours 55 minutes to 14 minutes 13 seconds, achieving a top-1 accuracy of 74.0%, which is $1.6\times$ and $2.9\times$ faster than previous 3D video models with higher accuracy. The code and more details can be found here: <http://tsm-hanlab.mit.edu>.

1 Introduction

Videos are increasing explosively in various areas, including healthcare, self-driving cars, virtual reality, *etc.* To handle the massive amount of videos collected everyday, we need a scalable video training system to enable fast learning. Existing works on distributed training mostly focus on image recognition [5, 21, 23, 2, 15, 22, 20]. For video recognition, the problem is more challenging but less explored: (1) the video models usually consume an order of magnitude larger computation compared to the 2D image counterpart. For example, a widely used ResNet-50 [7] model has around 4G FLOPs, while a ResNet-50 I3D [19] consumes 33G FLOPs, more than $8\times$ larger; (2) video datasets are far more larger than 2D image datasets, and the data I/O is much higher than images. For example, ImageNet [12] has 1.28M training images, while a video dataset Kinetics-400 has 63M training frames, which is around $50\times$ larger; (3) video models are usually larger in model size, thus it requires higher networking bandwidth to exchange the gradients.

In this paper, we study the bottlenecks of large-scale distributed training on videos, including computation, data loading (I/O), and communication. Correspondingly, we propose three practical design guidelines to solve the challenges: the model should leverage hardware-friendly operators to reduce the computation FLOPs; the model should take fewer input frames to save the file system I/O; the model should use operators with fewer parameters to save the networking bandwidth. Under such

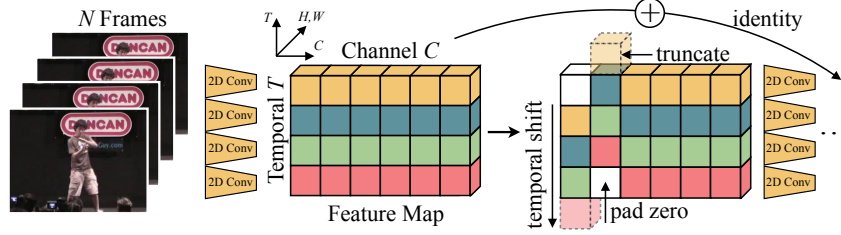


Figure 1. Temporal Shift Module (TSM) shifts channels along temporal dimension to enable temporal modeling among neighboring frames. It has *zero FLOPs* and *zero parameters*. It can be inserted into 2D CNNs to enable spatial-temporal feature learning.

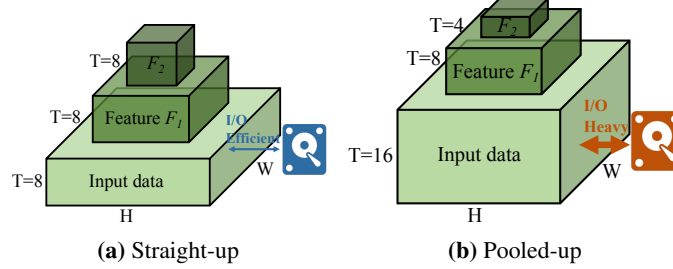


Figure 2. Two kinds of video backbone design. Straight-up backbone does not perform temporal pooling and is more data efficient. Pooled-up version requires many input frames and drains I/O.

guidelines, we propose an efficient video CNN operator: the Temporal Shift Module (TSM) that has *zero FLOPs* and *zero parameters*. It can scale up the Kinetics training to 1,536 GPUs, reaching a mini-batch size of 12,288 video clips/98,304 images. The whole training process can finish within 15 minutes and achieve a top-1 accuracy of 74.0%. Compared to previous I3D models [6, 19], our TSM model can increase the training speed by $1.6\times$ and $2.9\times$ on the world-leading supercomputer Summit.

2 Preliminary

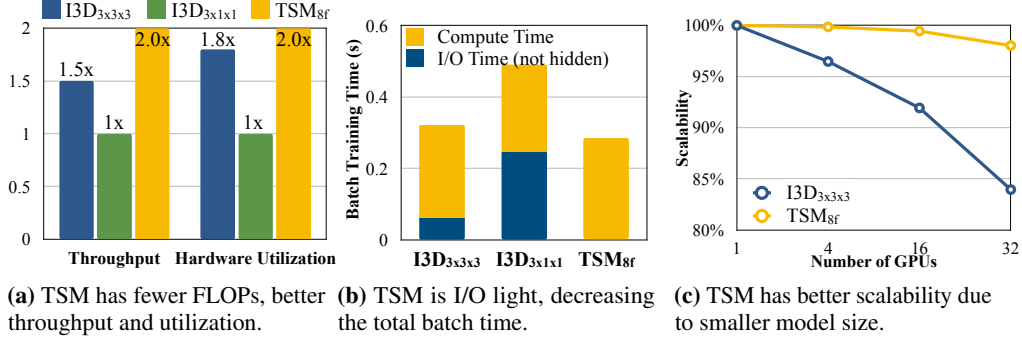
We first introduce the preliminary knowledge about video recognition networks. Different from 2D image recognition, the input/activation of a video network has 5 dimensions $[N, T, C, H, W]$, where N is the batch size, T is the temporal timestamps, C is the channel number, H and W are the spatial resolution. A simple video recognition model directly applies 2D CNN to each frames [18, 9, 14]. However, such methods cannot model the temporal relationships between frames, which are necessary for video understanding [11, 19]. To perform spatial-temporal feature learning, an effective way is to inflate 2D convolutions into 3D convolutions. The resulting model is I3D [1, 16]. However, by inflating temporal dimension by k , the computation and the model size is also increased by k times. Given such dilemma, we study two aspects of model design to make it more hardware efficient:

Temporal modeling unit. 3D convolution is the most widely used operator for spatial-temporal modeling. However, it suffers from two problems: (1) large computation and large parameter size, which slows down training and communication; (2) low hardware efficiency compared to 2D convolution (see Section 3 for details). 2D convolution kernels are highly optimized on various hardware and systems [4, 3], but 3D convolutions are less optimized. Give the same amount of FLOPs, 3D kernels run 1.2 to 3 times slower than 2D on cuDNN [4]. To deal with these issues, a more efficient operator is the Temporal Shift Module (TSM) [11]. As shown in Figure 1, TSM module shifts part of the channels along temporal dimension to mingle the feature between consecutive frames. In this way, we can perform temporal feature learning without incurring extra computation cost. With TSM, we can build a spatial-temporal video network using pure 2D CNN, enjoying high hardware efficiency.

Backbone topology. Existing video networks usually sample **many** frames as input (32 frames [19] or 64 frames [1]), and perform temporal pooling later to progressively *reduce* the temporal resolution (Figure 2b). Another way is to sample **fewer** frames (e.g. 8 frames [11]) as input while keeping the *same* temporal resolution to keep the information (Figure 2a). Although the overall computation of

Table 1. Efficiency statistics of different models. Arrows show the better direction.

| | Acc.↑ | FLOPs↓ | #Param.↓ | Input size↓ | Throughput↑ | Compute/IO↑ |
|---------------------------|--------------|--------|----------|-----------------------|-------------------------|---------------------|
| I3D _{3×3×3} [6] | 68.0% | 40G | 47.0M | 16×3×224 ² | 63.1V/s (1.5×) | 16.6k (2.4×) |
| I3D _{3×1×1} [19] | 73.3% | 33G | 29.3M | 32×3×224 ² | 41.9V/s (1.0×) | 6.85k (1×) |
| TSM [11] | 74.1% | 33G | 24.3M | 8×3×224 ² | 84.8V/s (2.0×) | 27.4k (4×) |

**Figure 3.** Analyzing how different design aspects influence the distributed training scalability of video recognition models: (a) **computation efficiency**; (b) **data loading efficiency**; (c) **networking efficiency**.

the two designs are similar, the former significantly increases the data loading traffic, making the system I/O heavy, which could be quite challenging in a distributed system considering the limited disk bandwidth.

3 Design Guidelines to Video Model Architecture

To tackle the challenge in a distributed training systems, we propose three video model design guidelines: (1) To increase the *computation efficiency*, use operators with lower FLOPs and higher hardware efficiency; (2) To reduce *data loading traffic*, use a network topology with higher FLOPs/data ratio; (3) To reduce the *networking traffic*, use operators with fewer parameters.

We show the advantage of the above three design guidelines by experimenting on three models in Table 1. All the models use the ResNet-50 backbone to exclude the influence of spatial modeling. The model architectures are introduced as follows.

(1) The first model is an I3D model from [6]. The model takes 16 frames as input and inflate all the 3×3 convolutions to $3 \times 3 \times 3$. It performs temporal dimension pooling by four times to reduce the temporal resolution. We denote the model as I3D_{3×3×3}.

(2) The second model is an I3D model from [19], taking 32 frames as input and inflating the first 1×1 convolution in every other ResBlock. It applies temporal dimension pooling by three times. We denote this more computation and parameter efficient design as I3D_{3×1×1}.

(3) The third model is built with TSM [11]. The TSM operator is inserted into every ResBlock. The model takes 8 frames as input and performs no temporal pooling. We denote this model as TSM.

a. Computation Efficiency. Computation efficiency is the most direct factor that influence the training time. As shown in Table 1, TSM_{8f} has $1.2\times$ fewer FLOPs compared to I3D_{3×3×3} and roughly the same FLOPs compared to I3D_{3×1×1}. However, the actual inference throughput also depends on the hardware utilization. We measure the inference throughput (defined as videos per second) of the three models on a single NVIDIA Tesla P100 GPU using batch size 16. We also measured the hardware utilization, defined as achieved FLOPs/second over peak FLOPs/second. The inference throughput and the hardware efficiency comparison is shown in Figure 3a. We can find that the *model is more hardware efficient if it has more 2D convolutions than 3D*: TSM is a fully 2D CNN, therefore it has the best hardware utilization ($2.0\times$); while the last several stage of I3D_{3×3×3} (res₄, res₅) have few temporal resolution (as shown in Table 2), it is more similar to 2D convolution and thus is $1.8\times$ more hardware efficient than I3D_{3×1×1} ($1.0\times$).

Table 2. The temporal resolution of output feature map for each block. TSM is a fully 2D structure, enjoying the best hardware efficiency. The last several stages of I3D_{3×3×3} have fewer temporal resolution, making it more similar to 2D CNN, thus enjoying better hardware efficiency compared to I3D_{3×1×1}.

| Block | data | conv ₁ | pool ₁ | res ₂ | res ₃ | res ₃ | res ₄ | res ₅ | global pool |
|----------------------|------|-------------------|-------------------|------------------|------------------|------------------|------------------|------------------|-------------|
| I3D _{3×3×3} | 16 | 16 | 8 | 8 | - | 4 | 2 | 1 | 1 |
| I3D _{3×1×1} | 32 | 16 | 8 | 8 | 4 | 4 | 4 | 4 | 1 |
| TSM _{8f} | 8 | 8 | 8 | 8 | - | 8 | 8 | 8 | 1 |

b. Data Loading Efficiency. Video datasets are usually very large. For a distributed system like the Summit supercomputer, the data is usually stored in High Performance Storage System (HPSS) shared across all the worker nodes. Such file systems usually have great sequential I/O performance but inferior random access performance. Therefore, large data traffic could easily become the system bottleneck. Previous popular I3D models [1, 6] takes many frames per video (16 or 32) as input and perform down-sample over temporal dimension. We argue that such design is a waste of disk bandwidth: a TSM_{8f} only takes 8 frames as input while achieving better accuracy. The intuition is that nearby frames are similar; loading too many similar frames is redundant. We empirically test the data loading bottleneck on Summit. To exclude the communication cost from the experiments, we perform timing on single-node training. We measure the total time of one-batch training and the time for data loading (that is not hidden by the computation). As shown in Figure 3b, for I3D_{3×1×1}, it takes 32 frames as input. The data loading time cannot be hidden by the computation, therefore data I/O becomes the bottleneck. I3D_{3×3×3} that takes 16 frame as input has less problem on data loading, while TSM_{8f} can fully hide the data loading time with computation. We also compute the model FLOPs divided by the input data size as a measurement of data efficiency. The value is denoted as "Compute/IO" as in Table 1. For scalable video recognition models, we want a model with larger Compute/IO ratio.

c. Networking Efficiency. In distributed training system, the communication time can be modelled as:

$$\text{communication time} = \text{latency} + \frac{\text{model size}}{\text{bandwidth}} \quad (1)$$

The latency and bandwidth is determined by the network condition, which cannot be optimized through model design. However, we can reduce the model size to reduce the communication cost. Both I3D models inflate some of the 2D convolution kernels to 3D, which will increase the number of parameters by k_T . While TSM module does not introduce extra parameters. Therefore, it has the same model size as the 2D counterpart. For example, I3D_{3×3×3} has $1.9\times$ larger model size than TSM_{8f}, which introduces almost two times of network communication during distributed training. To test the influence of model size on scalability, we measure the scalability on a 8 node cluster. Each computer has 4 NVIDIA TESLA P100 GPUs. We define the scalability as the actual training speed divided by the ideal training speed (single machine training speed * number of nodes). The results are shown in Figure 3c. Even with the high-speed connection, the scalability of I3D_{3×3×3} quickly drops as the number of training nodes increase: the scalability is smaller than 85% when applied to 8 nodes. While TSM_{8f} model still has over 98% of scalability thanks to the smaller model size thus smaller networking traffic.

4 Large-scale Distributed Training on Summit

In this section, we scale up the training of video recognition model on Summit supercomputer. With the help of above hardware-aware model design techniques, we can scale up the training to 1536 GPUs, finishing the training of Kinetics in 15 minutes.

4.1 Setups

Summit supercomputer. Summit [17] or OLCF-4 is a supercomputer at Oak Ridge National Laboratory, which as of September 2019 is the fastest supercomputer in the world. It consists of approximately 4,600 compute nodes, each with two IBM POWER9 processors and six NVIDIA Volta V100 accelerators. The POWER9 processor is connected via dual NVLINK bricks, each capable of

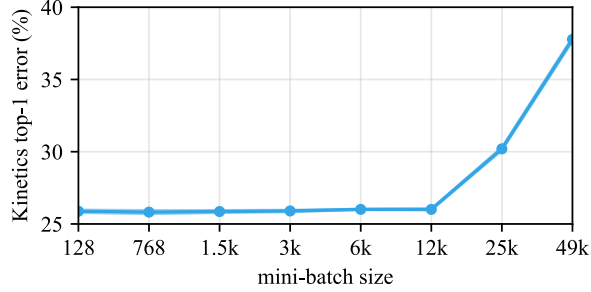


Figure 4. Kinetics top-1 validation accuracy vs. mini-batch size. The performance of the model does not degrade when we scale up the mini-batch size to 12k. The mean and standard deviation (the scale of the STD is hardly visible) are shown in the figure.

Table 3. Detailed statistics of different mini-batch size (* indicates simulated performance).

| #Node | #GPU | BatchSize | #Frames | Accuracy | Train Time | Note |
|-------|------|-----------|---------|------------|------------|------------------------------|
| 1 | 6 | 96 | 768 | 74.12±0.11 | 49h 55m | Baseline |
| 8 | 48 | 384 | 3,072 | 74.12±0.08 | 7h 7m | same level of accuracy |
| 16 | 96 | 768 | 6,144 | 74.18±0.14 | 3h 38m | |
| 32 | 192 | 1,536 | 12,288 | 74.14±0.10 | 1h 50m | |
| 64 | 384 | 3,072 | 24,576 | 74.10±0.08 | 55m 56s | |
| 128 | 768 | 6,144 | 49,152 | 73.99±0.04 | 28m 14s | |
| 256 | 1536 | 12,288 | 98,304 | 73.99±0.07 | 14m 13s | lose accuracy |
| 384 | 2304 | 18,432 | 147,456 | 72.52±0.07 | 10m 9s | |
| 512* | 3072 | 24,576 | 196,608 | 69.80±0.13 | - | |
| 1024* | 6144 | 49,152 | 393,216 | 62.22±0.17 | - | |

a 25GB/s transfer rate in each direction. Nodes contain 512 GB of DDR4 memory for use by the POWER9 processors and 96 GB of High Bandwidth Memory (HBM2) for use by the accelerators ¹.

We used PyTorch and Horovod [13] for distributed training. The framework uses ring-allreduce algorithm to perform synchronized SGD. The training is accelerated by CUDA and cuDNN. We used NVIDIA Collective Communication Library (NCCL) ² for most of the communication.

Dataset. In our experiments, we used Kinetics-400 dataset [10]. The dataset contains 400 human action classes, with at least 400 videos per classes. It contains human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands and hugging. It is widely used to benchmark the performance of video recognition models.

The dataset has roughly 240k training videos and 20k validation videos, each lasts around 10 seconds. Such large scale raises a critical challenge for model training and data storage.

Training/Testing Setting. We follow the commonly used distributed training and testing setting in [19]. For training, we train the network for 100 epochs. We denote the number of total GPUs used for training as k , and the batch size per GPU is denoted as n . The total batch size is kn . In our experiments, we trained a TSM network with 8-frame input and used a fixed $n = 8$. The initial learning rate is set to 0.00125 for every 8 samples, and we apply linear scaling rule [5] to scale up the learning rate with larger batch size. The total learning rate is $\eta = 0.00125k$. We used cosine learning rate decay with 5 epochs of warm-up [5]. We used weight decay $1e-4$ and no dropout for training. We applied no weight decay on BatchNorm and bias following [18, 8].

For testing, we follow [19] to sample 10 clips per video and calculate the average prediction. The video is spatially resized with a shorter size 256 and fed into network. For the reported results, we calculate the average and standard deviation from the checkpoint of last 5 epochs to reduce the influence of randomness,

¹<https://www.olcf.ornl.gov/for-users/system-user-guides/summit/summit-user-guide>

²<https://developer.nvidia.com/nccl>

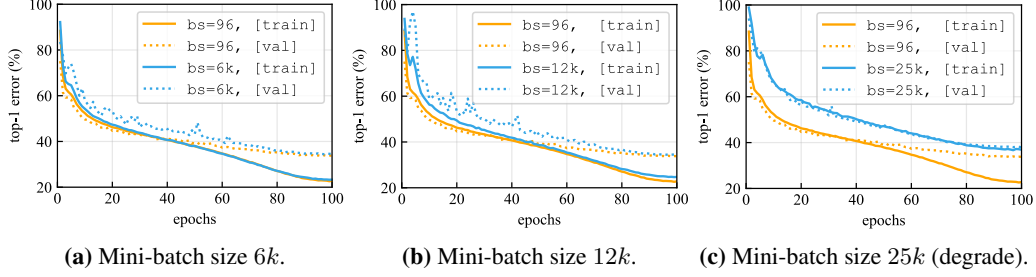


Figure 5. The learning curve for baseline training and large-batch distributed training (batch size 6144, 12228, 24576). The performance does not degrade for batch size 6144 and 12228, while degrades for a batch size of 24576.

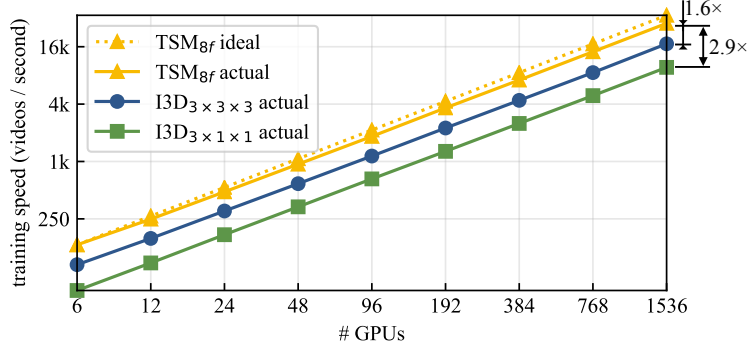


Figure 6. The training speed and scalability of distributed synchronous SGD training. TSM_{8f} achieves a good scalability ($>80\%$) even when using 1536 GPUs. TSM_{8f} can achieve $1.6\times$ higher training speed compared to I3D_{3 \times 3 \times 3} and $2.9\times$ compared to I3D_{3 \times 1 \times 1}, showing the effectiveness of the proposed design guidelines.

4.2 Experiments

Baseline. For the baseline, we trained a ResNet-50 TSM_{8f} model on a single Summit node with 6 GPUs, each GPU contains 16 video clips, resulting in a total batch size of $kn = 96$. We evaluate the performance of last 5 checkpoints, it achieves a top-1 accuracy of $74.12 \pm 0.11\%$.

Performance vs. Batch Size. We first compare the training error vs. the batch size. As shown in [5], the accuracy will not degrade when the batch size is relatively small. Therefore, our experiments start from 8 computing nodes (48 GPUs, 384 video clips, 3072 frames) to 1024 computing nodes (6144 GPUs, 49152 video clips, 393216 frames). Note that each sample in a video recognition model is a video clip consisting of several frames/images (in our case, 8). Therefore, the actual number of images used in one batch could be much larger than ImageNet training (*e.g.*, $98k$ vs. $8k$ [5]).

We first plot the error vs. batch size trade-off in Figure 4. The error does not increase when we scale the number of computing nodes up to 256 (1536 GPU), where the batch size is 12288, the total frame number is 98304. The detailed statistics are shown in Table 3. The scalability of TSM model is very close to the ideal case. Note that due to quota limitation, the largest physical nodes we used is 384 with 2304 GPUs. For 512 and 1024 nodes, we used gradient accumulation to simulate the training process (denoted by *).

We also provide the training and testing convergence curves using 768, 1536 and 3072 GPUs in Figure 5. For 768 GPUs and 1536 GPUs, although the convergence of large-batch distributed training is slower than single-machine training baseline, the final converged accuracy is similar, so that the model does not lose accuracy. For 3072 GPUs, the accuracy degrades for both training and testing.

Scalability We test the scalability of distributed training on Summit. According to the results from last section, we can keep the accuracy all the way to 256 computing nodes. Therefore, we sweep the number of computing nodes from 1 to 256 to measure the scalability. We keep a batch size of 8 for each GPU and each node has 6 GPUs. So the batch sizes change from 48 to 18,432. Each video clips contains 8 frames in our model, resulting a total number of frames from 384 to 147,456. We measure the training speed (videos/second) to get the actual speed-up. We calculate the ideal training speed using the single node training speed multiplied by number of nodes. The comparison

of different models is provided in Figure 6. The actual training speed is just marginally below the ideal scaling, achieving $> 80\%$ scaling efficiency. We also provide the detailed overall training time in Table 3. With 1536 GPUs, we can finish the Kinetics training with TSM within 14 minutes and 13 seconds, achieving a top-1 accuracy of 74.0%. The overall training speed of TSM_{8f} is $1.6\times$ larger than I3D_{3 \times 3 \times 3} and $2.9\times$ larger than I3D_{3 \times 1 \times 1}, showing the advantage of hardware-aware model design.

5 Conclusion

In this paper, we analyze how the design of video recognition models will affect the distributed training scalability. We propose three design guidelines to increase *computation efficiency*, *data loading efficiency*, and *networking efficiency*. With these guidelines, we designed a special TSM model and scale up the training on Summit supercomputer to 1536 GPUs, achieving 74.0% top-1 accuracy on Kinetics dataset within 14 minutes and 13 seconds.

Acknowledgments We thank John Cohn, Oak Ridge National Lab, MIT Quest for Intelligence, MIT-IBM Watson AI Lab, and AWS for supporting this work.

References

- [1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE, 2017.
- [2] Jianmin Chen, Xinghao Pan, Rajat Monga, Samy Bengio, and Rafal Jozefowicz. Revisiting distributed synchronous sgd. *arXiv preprint arXiv:1604.00981*, 2016.
- [3] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Haichen Shen, Meghan Cowan, Leyuan Wang, Yuwei Hu, Luis Ceze, et al. {TVM}: An automated end-to-end optimizing compiler for deep learning. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*, pages 578–594, 2018.
- [4] Sharan Chetlur, Cliff Woolley, Philippe Vandermersch, Jonathan Cohen, John Tran, Bryan Catanzaro, and Evan Shelhamer. cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*, 2014.
- [5] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.
- [6] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.
- [7] 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.
- [8] Xianyan Jia, Shutao Song, Wei He, Yangzihao Wang, Haidong Rong, Feihu Zhou, Liqiang Xie, Zhenyu Guo, Yuanzhou Yang, Liwei Yu, et al. Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes. *arXiv preprint arXiv:1807.11205*, 2018.
- [9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.
- [10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- [11] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. *arXiv preprint arXiv:1811.08383*, 2018.
- [12] 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.

- [13] Alexander Sergeev and Mike Del Balso. Horovod: fast and easy distributed deep learning in tensorflow. *arXiv preprint arXiv:1802.05799*, 2018.
- [14] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.
- [15] Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. Don’t decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.
- [16] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
- [17] Sudharshan S Vazhkudai, Bronis R de Supinski, Arthur S Bland, Al Geist, James Sexton, Jim Kahle, Christopher J Zimmer, Scott Atchley, Sarp Oral, Don E Maxwell, et al. The design, deployment, and evaluation of the coral pre-exascale systems. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, page 52. IEEE Press, 2018.
- [18] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.
- [19] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *arXiv preprint arXiv:1711.07971*, 10, 2017.
- [20] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.
- [21] Yang You, Igor Gitman, and Boris Ginsburg. Scaling sgd batch size to 32k for imagenet training. *arXiv preprint arXiv:1708.03888*, 6, 2017.
- [22] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. 2019.
- [23] Yang You, Zhao Zhang, Cho-Jui Hsieh, James Demmel, and Kurt Keutzer. Imagenet training in minutes. In *Proceedings of the 47th International Conference on Parallel Processing*, page 1. ACM, 2018.