

FedSkel: Efficient Federated Learning on Heterogeneous Systems with Skeleton Gradients Update

Junyu Luo
luojunyu@buaa.edu.cn
SCSE, Beihang University

Jianlei Yang*
jianlei@buaa.edu.cn
SCSE, Beihang University

Xucheng Ye
yexucheng@kuaishou.com
Heterogeneous Computing Center,
Kuaishou Technology

Xin Guo
guoxin@act.buaa.edu.cn
SCSE, Beihang University

Weisheng Zhao
weisheng.zhao@buaa.edu.cn
SME, Beihang University

ABSTRACT

Federated learning aims to protect users' privacy while performing data analysis from different participants. However, it is challenging to guarantee the training efficiency on heterogeneous systems due to the various computational capabilities and communication bottlenecks. In this work, we propose FedSkel to enable computation-efficient and communication-efficient federated learning on edge devices by only updating the model's essential parts, named skeleton networks. FedSkel is evaluated on real edge devices with imbalanced datasets. Experimental results show that it could achieve up to 5.52× speedups for CONV layers' back-propagation, 1.82× speedups for the whole training process, and reduce 64.8% communication cost, with negligible accuracy loss.

CCS CONCEPTS

• **Computing methodologies** → **Distributed artificial intelligence**; Distributed computing methodologies; • **Human-centered computing** → *Ubiquitous and mobile computing*.

KEYWORDS

Federated learning, heterogeneous system, distributed learning

ACM Reference Format:

Junyu Luo, Jianlei Yang, Xucheng Ye, Xin Guo, and Weisheng Zhao. 2021. FedSkel: Efficient Federated Learning on Heterogeneous Systems with Skeleton Gradients Update. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21)*, November 1–5, 2021, Virtual Event, QLD, Australia. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3459637.3482107>

*Corresponding author is Jianlei Yang. Email: jianlei@buaa.edu.cn

This work is supported in part by the National Natural Science Foundation of China (Grant No. 62072019, 61602022), State Key Laboratory of Software Development Environment and the 111 Talent Program B16001. The source code of this paper is publicly available on: <https://github.com/BUAA-CF-Lab/FedSkel>.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8446-9/21/11...\$15.00

<https://doi.org/10.1145/3459637.3482107>

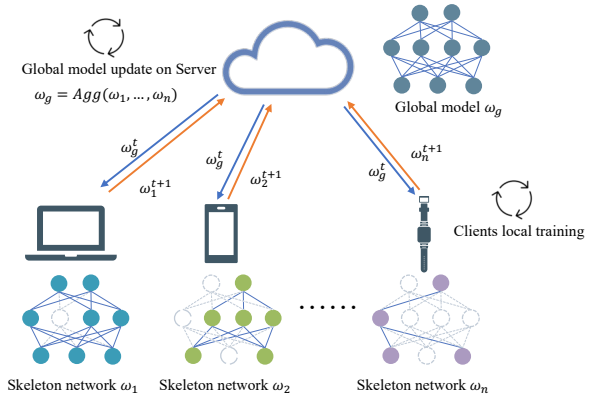


Figure 1: The framework of FedSkel. Each client selects its skeleton network and only updates (i.e. train, upload to server, download from server) the skeleton network.

1 INTRODUCTION

Federated learning (FL) [6] was proposed to protect data privacy while making use of the data collected from different participants. FL on edge devices or mobile devices utilizes multiple clients to collaboratively train the identical model on large amounts of local data. A global model can be trained by exchanging parameters between clients and servers instead of directly using private data. In practice, Google deployed FL to enhance models for emoji prediction [16] and query suggestions [20] applications.

Heterogeneities among clients will lead to inefficiency [13], which is a challenging issue when deploying FL systems. The heterogeneity caused by imbalanced computational capabilities [4] and communication bandwidth may significantly degrade the training speed. The global training process is limited by the slower clients, i.e., stragglers. If faster clients wait for slower ones, the overall training speed will become slow. Otherwise, the data in slower clients cannot be well utilized for global model learning.

Statistical heterogeneity is another challenge. In FL, data distribution across clients is inherently Non-IID (non-identically independently distributed). Thus, it is difficult for the shared global model to generalize for all clients. Some previous works [10, 12] tried to train personalized models on different clients. We also attempt to utilize data heterogeneity to make clients focus on different skeleton networks.

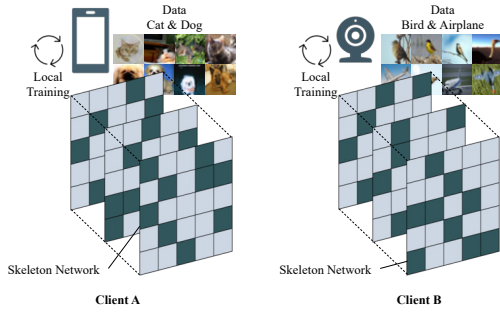


Figure 2: Different skeleton networks are existed in different clients since data distribution is imbalance in FL.

In this work, we propose FedSkel to improve FL efficiency on heterogeneous systems. As shown in Figure 1, each client determines its personalized skeleton network and it will only train and up/download the skeleton network. By adjusting the size of clients' skeleton networks, FedSkel can reduce workloads on slower clients and balance the latency across different clients to achieve efficient FL systems.

The main contributions of this paper are as follows:

- We introduce a novel federated learning framework, FedSkel, which achieves better computation efficiency on single clients and communication efficiency on FL systems.
- We propose to select personalized skeleton networks by a metric dynamically and perform gradients pruning to enable efficient training. We present the effectiveness of our method through experiments and analysis.
- We implement and measure speedups of FedSkel on real devices. We also compare FedSkel with three baselines to demonstrate it will not degrade accuracy while accelerating FL systems.

2 RELATED WORKS

Weight Pruning. Weight pruning usually aims to accelerate models inference by reducing the model parameters after training. [1, 8] exploit structural weight pruning. Yu [22] observed that some filters are more crucial than others on different categories of data. In this paper, we follow [22] to select personalized skeleton networks for clients. Different from weight pruning methods, we mainly focus on reducing training workloads.

Communication-Efficient Federated Learning. Extensive prior works attempt to speed up FL by decreasing parameter communications [6, 17, 19]. Lin [14] demonstrated that removing redundant gradients will not hurt accuracy. However, these works did not take efforts to reduce the computation of local devices. FedSkel can not only achieve better communication efficiency but also accelerate whole training process.

Personalization. Personalization is a critical challenge in federated learning. LG-FedAvg [12] and LotteryFL [10] train personalized models for each client. However, these methods did not consider the heterogeneity of computational capabilities in real FL systems. In this paper, FedSkel can facilitate efficient FL training with personalization maintained.

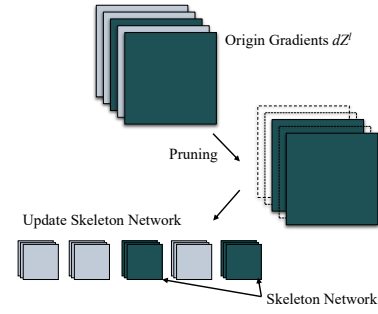


Figure 3: Structural gradients pruning for skeleton network update on clients training.

3 METHODOLOGIES

In FL, data imbalance is an inherent feature. Hence, submodels inherited from global model will be of different importance to different clients. From this motivation, we propose FedSkel. In FedSkel, each client only trains and up/downloads their important submodels, which are named skeleton networks. The central server aggregates all local skeleton networks to obtain the global model, which can handle tasks of different data distributions. Clients' training processes can be accelerated by only updating the skeleton networks.

3.1 Skeleton Network

Yu [22] found there are category-related filters in CNNs, which are more activated to the specific data categories and contribute more to the prediction. In FL scenario, as shown in Figure 2, clients' important filters (skeleton network) are different since they hold data of different distributions.

Importance Metric. Clients can select their skeleton network according to the following defined metric. The logits of CNNs can be calculated with l -th layer by (notice that we left out bias here):

$$Z(W, A^l) = (\alpha \dots \alpha (A^l W^{l+1}) \dots W^{L-1}) W^L, \quad (1)$$

$$M_i^l = |A_i^l|, \quad (2)$$

where the model is with total L layers, Z is the output, W^l and A^l is the weight and input to the l -th layer, α is the activation function.

As shown in Eq. 1 and Eq. 2, when W^{l+1} is non-zero, M_i^l determines the contribution of A_i^l to logits. Hence, we adopt M_i^l to measure the importance of i -th filter in l -th layer.

Gradients Pruning enables Efficient Training. Gradients pruning is utilized to only train skeleton networks and adjust workloads. CNNs training mainly includes three kinds of matrix multiplication operations, which take up most of the training computation [21]:

- *Forward:* $Z^l = A^{l-1} * W^l$
- *Gradients Back-Propagation:* $dA^{l-1} = dZ^l * W^l$
- *Weight Gradients Computation:* $dW^l = A^{l-1} * dZ^l$

where Z^l denotes the output of l -th layer. The last two multiplication operations are involved in back-propagation procedure. In this paper, we try to apply structured pruning on gradients dZ^l as Figure 3. Since we compress the dZ^l , the computation in *Gradients*

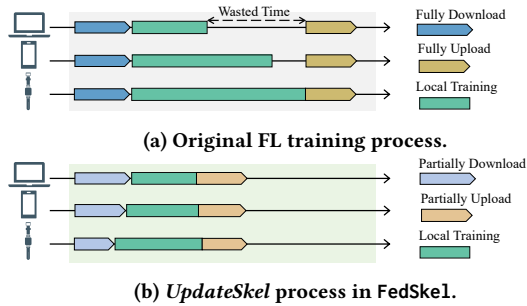


Figure 4: FedSkel reduces both local training cost and parameters communication cost. Partially up/download means FedSkel only needs to up/download skeleton networks.

Back-Propagation and *Weight Gradients Computation* can be greatly reduced, thus reducing the workloads on clients.

Gradients Pruning has Negligible Impact on Accuracy. [14] revealed the redundancy of gradients. Ye [21] exploited fine-grained gradient pruning for training acceleration. In this paper, structural gradients pruning is utilized since it is more hardware-friendly. FedProx [11] demonstrated that adding constraints to the updates from clients could benefit the issue of data heterogeneity. In FedSkel, we only update the submodels on clients, which also satisfy the constraints introduced by FedProx. Experiments in Section 4 show that FedSkel will not degrade accuracy and can even perform better than baselines.

3.2 FedSkel

FedSkel aims to solve the potential inefficiencies in heterogeneous FL systems. It tries to balance the workloads of different clients by setting different skeleton network ratios r .

The whole training procedure are divided into two alternately processes: *SetSkel* and *UpdateSkel*. In *SetSkel* process, skeleton networks are determined for each client. In *UpdateSkel* process, clients only update the skeleton network to reduce the training workloads. In practice, a *SetSkel* process is usually followed by 3 to 5 *UpdateSkel* processes.

SetSkel Process

The purpose of *SetSkel* is to select the skeleton network for each client. At the same time, let clients get updates related to non-skeleton networks by exchanging parameters with the server.

Server sets skeleton ratios r . In a n -clients system, the i -th client uploads its computational capability c_i to server. The server normalizes c as $c'_i = c_i/c_{max}$. We simply try to set skeleton ratios r with a linear function, and setting r more effectively can be further explored.

Calculating importance metric during training. The *SetSkel* process is similar to the standard FL process (as shown in Figure 4a). The only difference is that we accumulate the importance metric M_i^l and set the skeleton network according to them.

UpdateSkel Process

In *UpdateSkel* process, as shown in Figure 4b, clients reduce the computational workloads and communication workloads by only training and exchanging skeleton networks.

Table 1: Speedups on Intel CPU and ARM CPU with different skeleton ratio r .

r	Intel CPU		ARM CPU	
	Back-prop	Overall	Back-prop	Overall
40%	2.08×	1.10×	1.94×	1.35×
30%	2.57×	1.13×	3.06×	1.52×
20%	3.38×	1.21×	4.32×	1.61×
10%	5.52×	1.28×	4.56×	1.82×

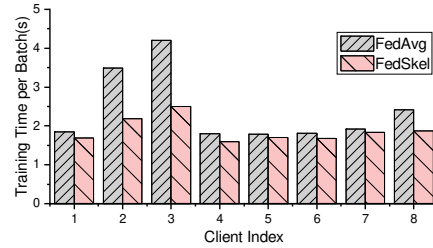


Figure 5: Runtime of each client for an 8-device system to train one batch with FedSkel and FedAvg, including forward and backward.

Computation reduction. In *UpdateSkel* Process, clients only train the skeleton network, the computational workloads can be reduced as described in Section 3.1.

Communication reduction. In *UpdateSkel* Process, clients only up/download parameters on skeleton network to/from the server. The amount of communication cost is determined by skeleton network ratio r .

Overall Procedure

SetSkel processes and *UpdateSkel* processes are iterated. In real world, *SetSkel* processes typically conduct in the period when computational resources are idle (such as in nights), so they do not take up lots of computational resources. The server adopts federated averaging [15] to aggregate updates to obtain the global model which can generalize to all type of data distribution.

Hence, FedSkel improves the learning efficiency by reducing both the computational workloads and communication cost.

4 EXPERIMENTS

In this section, we demonstrate that FedSkel is more training-efficient on heterogeneous systems without affecting accuracy.

4.1 Acceleration

In this part, we evaluate the speedups of FedSkel in *UpdateSkel* processes. We conduct experiments on a single device and a system of 8 devices, respectively.

Implementation. FedSkel is evaluated on Intel Xeon E5-2680 v4@2.4GHz CPU with MKL and Raspberry Pi 3B+ (ARM v8@1.4GHz) with OpenBLAS. We modify the back-propagation in Caffe's [5] CONV layer. Speedups are measured with LeNet [9] on MNIST [9]. Same as Section 3.2, r denotes the skeleton network ratio.

Acceleration on Single Device. To evaluate the performance with different r , we adopt $r = \{40\%, 30\%, 20\%, 10\%\}$. According to

Table 2: Volume of parameters communication for different methods with LeNet-5 on MNIST.

Method	Params Comm.	Reduction
FedAvg [15]	12.8×10^9	–
FedMTL [18]	12.0×10^9	6.3%
LG-FedAvg [12]	8.5×10^9	33.6%
FedSkel ($r = 10\%$)	4.5×10^9	64.8%

the results in Table 1, we can achieve up to $5.52\times$ speedups at back-propagation of CONV layers and $1.82\times$ at whole training process.

Acceleration on Edge Systems. We also deploy experiments on a real system with 8 Raspberry Pi. Devices are set to different computational capabilities c_i as in a heterogeneous system. We set r_i according to c_i . Figure 5 shows the time consuming in *UpdateSkel* process for one batch (batch_size=512) training. It shows that FedSkel can balance the workloads, speedup the slower clients and achieve up to $1.82\times$ speedups of the whole system.

4.2 Communication Cost Reduction

FedSkel can significantly reduce communication cost since we only exchange updates on the skeleton network in *UpdateSkel*. As shown in Table 2, FedSkel with $r = 10\%$ can reduce 64.8% communicate cost to the whole training process, including *SetSkel* and *UpdateSkel*.

4.3 Convergence and Accuracy

In this part, we empirically compare FedSkel with the baseline methods. We demonstrate that FedSkel can significantly accelerate training without hurting accuracy.

Experimental Settings

Datasets and Models. All methods are evaluated on MNIST [9], FEMNIST [2], CIFAR-10,100 [7] datasets and with LeNet-5 [9] and ResNet-18,34 [3]. We adopt the Non-IID data setting as [12] did. Each client is assigned with 2 shards of Non-IID splitted data for MNIST and CIFAR-10, while 20 shards for others. LeNet is trained for 1000 epochs and ResNets for 600 epochs. Each *SetSkel* process is followed by 3 *UpdateSkel* processes. To make a fair comparison, we follow the experimental design in [15] and also exploit local representation learning. The FL system consists of 1 central server and 100 clients. All methods are evaluated with the same settings.

Heterogeneous System Settings. In the real world, clients in FL systems are of different computational capabilities. To verify the performance of FedSkel in this scenario, we set each client with a different ratio r equidistant ranging from 10% to 100%.

Experimental Results

As shown in Table 3 and Table 4, crossing different datasets and models, our method will not hurt accuracy while achieving acceleration. FedSkel can even improve the personalization to achieve better accuracy on the local test.

[†]We follow the test settings of LG-FedAvg [12]. Local Test (new predictions on an existing device), where we test each client and the clients' train/test data are of the same distribution. New Test (new predictions on new devices), where we test the global model on test data with the same distribution of the whole dataset.

Table 3: Accuracy comparison of baselines and FedSkel (ours) on different datasets with LeNet.

Method	Test Type [†]	Dataset			
		MNIST	FEMNIST	CIFAR-10	CIFAR-100
FedAvg [15]	New	99.09	57.52	59.03	32.44
	Local	99.09	57.52	59.03	32.44
FedMTL [18]	New	39.76	24.69	10.32	2.15
	Local	99.41	29.93	90.49	42.68
LG-FedAvg [12]	New	99.09	57.57	58.48	32.44
	Local	99.45	78.92	92.47	52.95
FedSkel	New	99.09	59.43	58.48	32.52
	Local	99.46	82.95	92.60	53.66

Table 4: Accuracy comparison of baselines and FedSkel (ours) with different models on CIFAR-10 dataset.

Method	Test Type [†]	Model		
		LeNet	ResNet-18	ResNet-34
FedAvg [15]	New	59.03	67.61	70.28
	Local	59.03	67.61	70.28
FedMTL [18]	New	10.32	10.92	10.02
	Local	90.49	92.78	91.90
LG-FedAvg [12]	New	58.48	75.67	76.95
	Local	92.47	96.21	97.34
FedSkel	New	58.48	77.20	76.92
	Local	92.60	96.59	97.65

4.4 Analysis and Discussions

FedSkel will not affect accuracy though it prunes gradients in the training process. Several reasons are accounting for it.

- Skeleton network is the submodel which has a more crucial impact on predict results. Hence the combination of skeleton network is able to perform well on each task.
- FedSkel facilitates personalization by only updating skeleton networks. It enables clients to perform better on their own tasks.
- According to FedProx [11], fewer updates to the global model facilitate robust convergence. FedSkel only updates the skeleton network, also contributes to a faster and stable convergence.

Experiments show that FL is robust to gradients pruning. By optimizing the gradients' flow, we can achieve an efficient and personalized FL system.

5 CONCLUSIONS

In this work, we propose FedSkel as a new framework for heterogeneous FL systems. In our method, clients select skeleton networks and only update skeleton filters. Skeleton network ratios are adaptive to clients' computational capabilities. We have shown that our method does not affect the accuracy through extensive experiments and analysis, and achieves up to $5.52\times$ speedups in back-prop and $1.82\times$ speedups in the overall process on edge devices. Our approach enables efficient FL on heterogeneous systems. The future work can be the better metrics of selecting skeleton networks, strategies to set clients' skeleton ratios. We will also extend our explorations on the effect of gradient optimization in FL systems.

REFERENCES

- [1] Sajid Anwar, Kyuyeon Hwang, and Wonyong Sung. 2017. Structured pruning of deep convolutional neural networks. *ACM Journal on Emerging Technologies in Computing Systems* 13, 3 (2017), 32.
- [2] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. 2018. LEAF: A Benchmark for Federated Settings. *arXiv:1812.01097* [cs.LG]
- [3] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.
- [4] Andrey Ignatov, Radu Timofte, Andrei Kulik, Seungsoo Yang, Ke Wang, Felix Baum, Max Wu, Lirong Xu, and Luc Van Gool. 2019. AI Benchmark: All About Deep Learning on Smartphones in 2019. <https://ai-benchmark.com/ranking.html>.
- [5] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093* (2014).
- [6] Jakub Konečný, H. Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* (2016).
- [7] Alex Krizhevsky and Geoffrey Hinton. 2009. *Learning multiple layers of features from tiny images*. Technical Report. Citeseer.
- [8] Vadim Lebedev and Victor Lempitsky. 2016. Fast ConvNets using group-wise brain damage. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2554–2564.
- [9] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
- [10] Ang Li, Jingwei Sun, Binghui Wang, Lin Duan, Sicheng Li, Yiran Chen, and Hai Li. 2020. LotteryFL: Personalized and communication-efficient federated learning with lottery ticket hypothesis on Non-IID datasets. *arXiv preprint arXiv:2008.03371* (2020).
- [11] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2018. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127* (2018).
- [12] Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523* (2020).
- [13] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. 2020. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials* 22, 3 (2020), 2031–2063.
- [14] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. 2017. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887* (2017).
- [15] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. 2016. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629* (2016).
- [16] Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays. 2019. Federated learning for emoji prediction in a mobile keyboard. *arXiv preprint arXiv:1906.04329* (2019).
- [17] Shaohuai Shi, Qiang Wang, Kaiyong Zhao, Zhenheng Tang, Yuxin Wang, Xiang Huang, and Xiaowen Chu. 2019. A distributed synchronous SGD algorithm with global Top-k sparsification for low bandwidth networks. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2238–2247.
- [18] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. 2017. Federated multi-task learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 4427–4437.
- [19] Nikko Strom. 2015. Scalable distributed DNN training using commodity GPU cloud computing. In *Sixteenth Annual Conference of the International Speech Communication Association*.
- [20] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. 2018. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903* (2018).
- [21] Xucheng Ye, Pengcheng Dai, Junyu Luo, Xin Guo, Yingjie Qi, Jianlei Yang, and Yiran Chen. 2020. Accelerating CNN training by pruning activation gradients. In *European Conference on Computer Vision*. Springer, 322–338.
- [22] Fuxun Yu, Zhuwei Qin, and Xiang Chen. 2018. Distilling critical paths in convolutional neural networks. *arXiv preprint arXiv:1811.02643* (2018).