A APPENDIX

This appendix provides more details about the experiments and its results.

A.1 MORE EXPERIMENTAL RESULTS AND DETAILS

A.1.1 DETAILS OF THE SEARCH SPACE DEFINITION

We adopt the following 7 operations in all our experiments: 3×3 and 5×5 separable convolutions, 3×3 and 5×5 dilated separable convolutions, 3×3 max pooling, 3×3 average pooling, identity, and zero.

The network is formed by stacking convolutional cells multiple times. Cell k takes the outputs of cell $k-2$ and cell $k-1$ as its input. Each cell contains seven nodes: two input nodes, one output node, and four intermediate nodes inside the cell. The input of the first intermediate node is set equal to two input nodes, and the other intermediate nodes take all previous intermediate nodes' output as input. The output node concatenates all intermediate nodes' output depth-wise. There are two types of cells: the normal cell and the reduction cell. The reduction cell is designed to reduce the spatial resolution of feature maps located at 1/3 and 2/3 of the total depth of the network. Architecture parameters determine the discrete operation value between two nodes. All normal cells and all reduction cells share the same architecture parameters $\boldsymbol{\alpha}_{\mathrm{n}}$ and $\boldsymbol{\alpha}_{\mathrm{r}}$, respectively. By this definition, our method alternatively optimizes architecture parameters ($\boldsymbol{\alpha}_{\mathrm{n}}$, $\boldsymbol{\alpha}_{\mathrm{r}}$) and model weight parameters $\boldsymbol{w}$. Besides the search space, the other details of the system design can be found in our source code.

A.1.2 DETAILS OF THE HETEROGENEOUS DISTRIBUTION ON EACH CLIENT (NON-IID)

In this work, we performed experiments on CIFAR10 and gld23k datasets. For CIFAR10, we explored two types of non-IIDness, label-skewed and lda distribution. Table 3 shows the lda data distribution used in our experiment of global model search Via FedNAS. We can see that the sample number of each class in each worker is highly unbalanced. Some classes in a worker even have no samples, and some classes take up most of the proportion (highlighted in the table). For personalized experiments, we used two types of heterogeneity settings shown in Figures 4 and 7. As it can be seen that the distribution setting of 7 is challenging given not that the number of images per client varies but also the number of images belonging to a specific class.

Besides CIFAR10, we also evaluated personalized experiments on gld23k dataset. Since gld23k dataset have 203 clients data and some client's can have as low as 30 images and splitting it further in training and test dataset would make it insufficient for efficient training. Therefore, out of 203 clients, we only use those client's data which have images greater than 200. This condition would provide us sufficient data to perform search/training at each client and further test local inference to record client's validation accuracy. Figure 8 plots the image and label allocation per client for gld23k federated dataset under this setting. As it can be seen the distribution is non-IID especially in terms of label allocation per client.

A.1.3 RESULTS FOR CIFAR10 (LDA) AND GLD23K

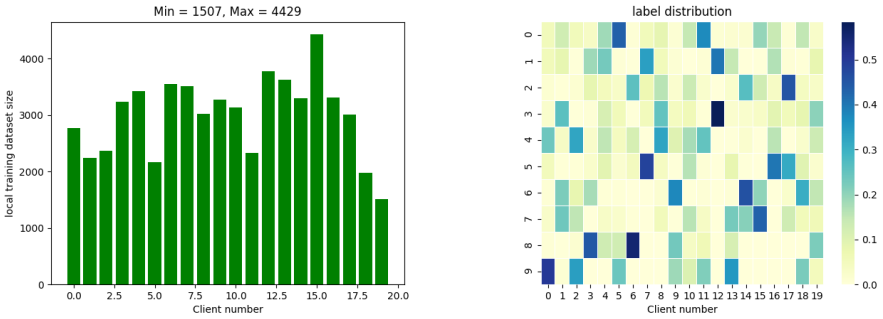
Figure 9 illustrates the results of comparison of `FedNAS` with `FedAvg` (with local adaptation), `perFedAvg`, `Ditto` and `FedNAS` with lda distribution of cifar10 (which is given in Figure 7). It can be seen that `FedNAS` outperforms all these methods. Since the number of rounds of convergence were for these methods, we plotted these figures separately for clarity. The best accuracy for this setting of `FedNAS` is 90.64% whereas `FedAvg` yields accuracy of 86.1%. On the other hand, we achieve 88.0% and 89.4% average validation accuracies of all the clients with `Ditto` and `perFedAvg`, respectively. Likewise, for gld23k we obtain 56.45%, , 45.28%, 43.92% and 34.5% accuracies with `FedNAS`, `Ditto`, `FedAVg with Local Adaptation` and `MAML`, respectively. The accuracy gap for gld23k between Ditto and FedNAS is more than 10%.

A.1.4 HYPERPARAMETER SETTING

We report important well-tuned hyperparameters used in our experiments. For global search experiments, FedNAS searches 50 communication rounds using five local searching epochs, with a batch

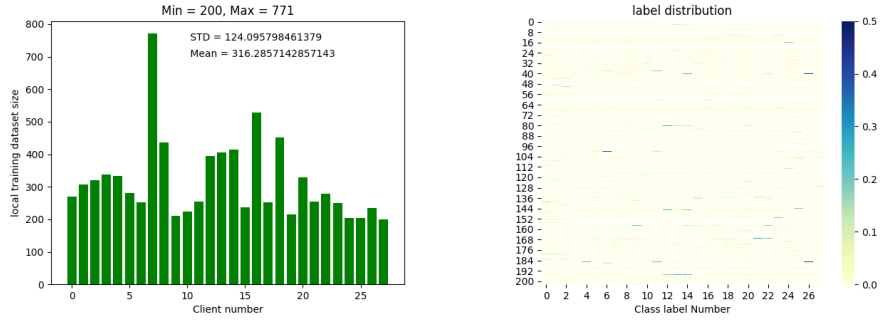
Client ID	Numbers of samples in the classes										Distribution
	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	
k=0	144	94	**1561**	133	**1099**	**1466**	**0**	**0**	**0**	**0**	
k=1	327	28	264	16	354	2	100	20	200	3	
k=2	6	6	641	1	255	4	1	2	106	**1723**	
k=3	176	792	100	28	76	508	991	416	215	**0**	
k=4	84	**1926**	1	408	133	24	771	**0**	**0**	**0**	
k=5	41	46	377	541	7	235	54	**1687**	666	**0**	
k=6	134	181	505	720	123	210	44	58	663	221	
k=7	87	2	131	**1325**	**1117**	704	**0**	**0**	**0**	**0**	
k=8	178	101	5	32	**1553**	10	163	9	437	131	
k=9	94	125	**0**	147	287	100	23	217	608	279	
k=10	379	649	106	90	35	119	807	819	3	85	
k=11	**1306**	55	681	227	202	34	**0**	648	**0**	**0**	
k=12	**1045**	13	53	6	77	70	482	7	761	494	
k=13	731	883	15	161	387	552	4	**1051**	**0**	**0**	
k=14	4	97	467	899	**0**	407	50	64	**1098**	797	
k=15	264	2	93	266	412	142	806	2	243	**1267**	

Table 3: Heterogeneous data distribution (non-IID) used in FedNAS for Global Model experiments



(a) Image Allocation per Client (b) Label Allocation per Client

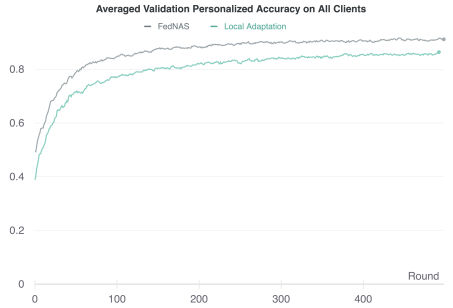
Figure 7: Heterogeneous data distribution (non-IID) used in FedNAS for Personalized Model experiments



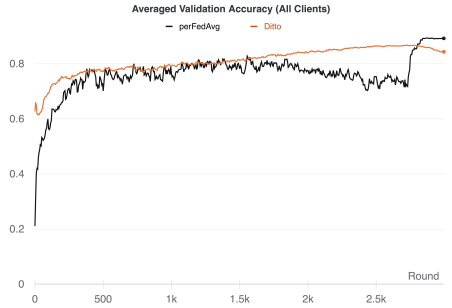
(a) Image Allocation per Client (b) Label Allocation per Client

Figure 8: Heterogeneous data distribution (non-IID) with Federated gld23k used in FedNAS for Personalized Model experiments

size of 64. For FedAvg, DenseNet201 is used for training, with 100 communication rounds, 20 local epochs, a learning rate of 0.08, and a batch size of 64. Both methods use the same data augmentation



(a) Average Validation Accuracy all clients for FedNAS and Local Adaptation



(b) Average Validation Accuracy Clients for Ditto and PerFedAvg

Figure 9: Average Validation Accuracy for CIFAR10 LDA Partition ($\alpha = 0.5$)

techniques that are used in image classification, such as random crops, flips, and normalization. More details and other parameter settings can be found in our source code.

For personalized models search, we explored CIFAR10 with label skewed and lda distribution. For label skew distribution, we searched over 500 communication rounds with batch size 32 for FedNAS and Local Adaptation, and 3000 rounds for Ditto and perFedAvg. We searched hyperparameters over the lr set of {0.1,0.3, 0.01, 0.03, 0.003, 0.001} and found the best lr to be 0.01, 0.01, 0.001, 0.003 for `FedAvg` with local adaption, `FedNAS`, `Ditto`, `perFedAvg`, respectively, for both label skewed and lda distribution with cifar10. For gld23k, we searched over the same hyperparameters and found the best performing hyperparameters to be 0.1, 0.1, 0.001, 0.003 for `FedAvg` with local adaption, `FedNAS`, `Ditto`, `perFedAvg`, respectively.

A.1.5 Visualization of the Search Architecture

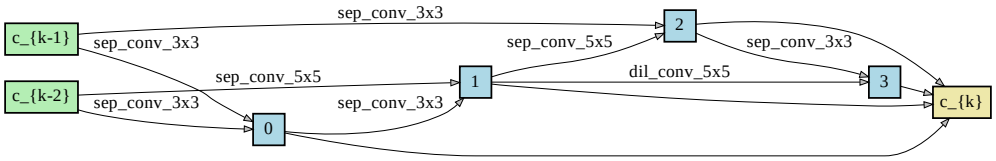


Figure 10: Normal Cell Architecture

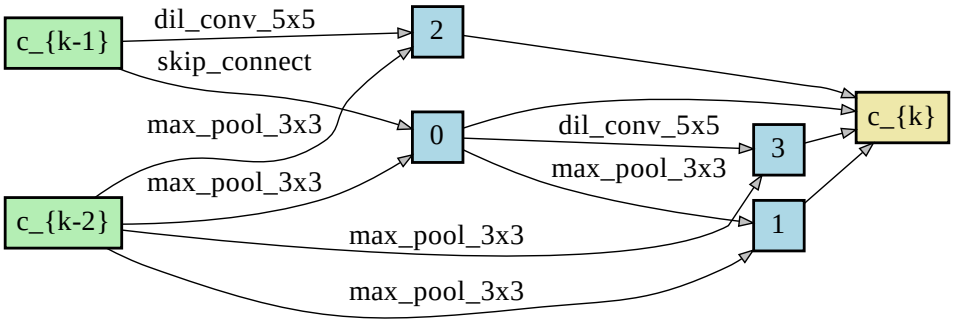


Figure 11: Reduction Cell Architecture

We report the architecture searched based on the above given non-IID dataset and hyper-parameter setting for global experiments of FedNAS. Figures 10 and 11 show the normal cell architecture and the reduction cell architecture, respectively. We can see that the reduction cell uses more pooling operations while the normal cell has more convolutional operations.

B FUTURE WORKS

Our future work aims to improve the FedNAS framework from form the following perspectives.

- **Local NAS under Resource Constraint**. Our current search space fits cross-organization federated learning, where the edge device can be equipped with powerful GPU devices. But when used in resource-constrained environments such as smartphones or IoT devices, the memory of our search space is too large. Searching on compact search space or using sampling methods are potential solutions to this challenge.
- **Privacy-preserved FedNAS**. In this work, we explored neural architecture search where client communicates its architecture parameters α and model parameters w with the server. We also showed that FedNAS has the potential to yield personalization benefits. Given this context, revealing both α and w to adversary may provide more information than sharing only w to server with a predetermined model. Therefore, exploration of privacy preserved FedNAS can be an interesting but a challenging direction to investigate.
- **Transferability and Federated Learning with Weight Sharing**. Another interesting direction would be transferring the searched architectures on each client. It is important to note that after the transfer, each client may have a different architecture, therefore, conventional FL weight aggregation may not work. To train this transferred models, one can explore weight sharing to train these models in federated setting.