Dynamic Channel Pruning for Privacy

Abhishek Singh
MIT

Vivek Sharma
MIT

Ayush Chopra
MIT

Ethan Z Garza
MIT

Praneeth Vepakomma
MIT

Ramesh Raskar
MIT

Abstract

Increased deployment of deep learning services on remote cloud instances requires users to share their sensitive information with un-trusted parties. Correspondingly, we focus on private inference in these distributed learning setup via sharing of intermediate activations instead of raw inputs. Specifically, we design a dynamic pruning strategy to mitigate information leakage from activations and empirically validate effectiveness of the proposed technique. We analyze performance and discuss ethical considerations for deploying allied services for general availability.

1 Introduction

With increasing deployment of deep learning applications on the cloud, users are required to share their sensitive data with parties that they do not necessarily trust. We propose a method where individuals can perform inference (prediction) tasks by sharing activations from an intermediate layer instead of the raw sensitive data. However, a dishonest entity (server) may attempt to utilize these intermediate activations to reconstruct the input or predict sensitive attributes. We design a dynamic pruning strategy to mitigate information leakage from activations. Channel pruning in deep network compression posits that there is significant representational redundancy in learned activations [2, 15, 22, 7] and seeks to eliminate it in order to reduce computational complexity and accelerate inference. We hypothesize that privacy sensitive information in a given image is a result of underlying human interpretable features and pruning corresponding channels can help suppress leakage of sensitive information. Specifically, we focus on mitigating attribute inference attacks where a dishonest server (or *adversary*) intends to extract private characteristics (*sensitive attributes*) of the raw input while also preserving utility which is defined by accuracy on a main task (on a *prediction attribute*). Correspondingly, we propose a dynamic channel pruning technique mechanism to selectively hide sensitive attributes and achieve a privacy-utility trade-off by adjusting the number of channels to be pruned. We show the efficacy of our proposed mechanism by analyzing the performance of the main task and a simulated adversary. The simulated adversary tries to mimic a worst-case attacker who uses a subset of intermediate activations for obtaining sensitive information.

Threat Model The first step to address any privacy or security concern is to perform the threat modeling for the system. In our setup, the deep learning model ($h(\theta;)$) is distributed across client and server, client generates the above-mentioned intermediate activations (z) from the raw input (x), and server processes these activations z for the prediction. These client and server models are denoted by $f(\theta_1;)$ and $g(\theta_2;)$ respectively such that $h(\theta;x) = f(\theta_1; g(\theta_2;x))$. To protect the raw input x, the client communicates $z = g(\theta_2;x)$ to the server which then generates the output prediction $y_T = f(\theta_1;z)$ for a main task T. While y_T is information that the server can visualize and process, an adversary may additional seek to leverage z to extract a private attribute y_p of the input x. Depending upon the formulation of y_p, the attack by the adversary can be: a) *reconstruction attack* where goal is estimate the input (i.e. y_p is x) b) *attribute inference attack* where y_p is an sensitive

Preprint. Under review.

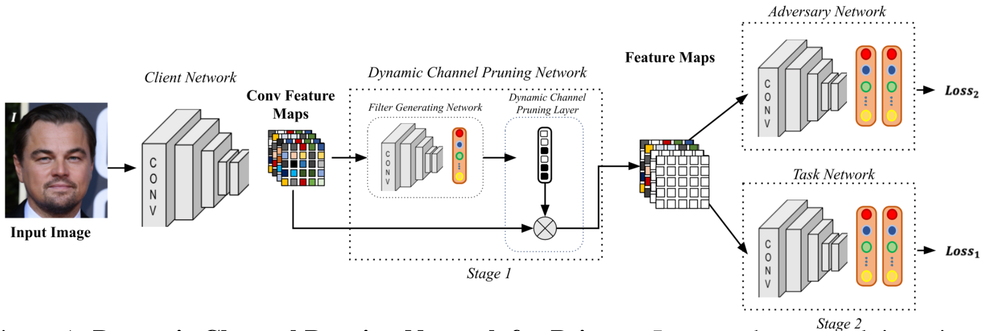


Figure 1: **Dynamic Channel Pruning Network for Privacy**. Input to the network is an image, as well as task labels and attribute labels to hide. The network is jointly optimized with a task objective to adaptively hide a given attribute without drop in performance of the target task.

attribute associated with x that the client intends to keep private. Under this threat model, information leakage attacks can manifest in three ways:

- Learning based attacks - In the learning based methods, the attacker trains a separate adversary model $f_A(\theta_3;)$ by obtaining a set of publicly available or leaked records of (x, z) pairs. In computer vision literature, this is also referred as *expected pre-image* [5]
- Optimization based attacks - In optimization based approaches, the weights θ_2 are known to the adversary. The adversary starts out with a random image x_0 and optimizes it to obtain $x^* = \arg\min_{x_0} \ell(g(\theta_2; x_0), z)$. Additional regulariser like total variance [12] terms have shown to improve performance of the reconstructed image.
- Prior based attacks - This aligns with the work done in image restoration tasks using priors like deep image prior [19]. This attack scheme is the strongest of all because it does not assume any prior knowledge about leaked subset or knowledge of the weights of the model.

2 Related Work

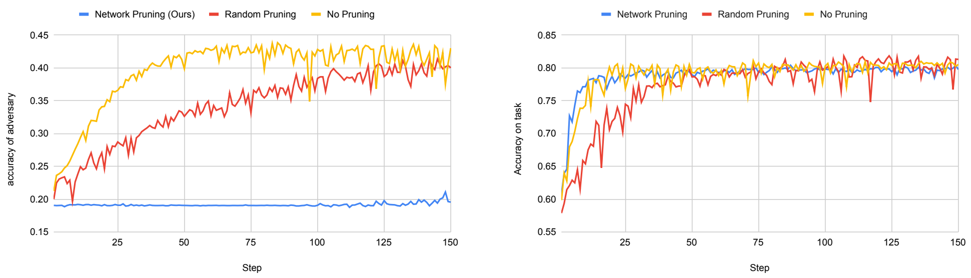
Privacy-preserving learning has majorly focused on privacy of training data, especially in distributed leaning settings. The key techniques for distributed training include federated learning [13], where entire models are trained on local data at client, and split learning [8, 20], where models are *split* between client and server with communication via intermediate activations. To preserve privacy, various mechanisms have been proposed for secure aggregation [3] and to reduce information leakage from (trained) model parameters [1, 18]. While majority of the work has been done for privacy of the training data, *focus of this work is privacy of data in distributed inference*.

Private inference Osia et al. [16, 17] propose sharing extracted features and give upper and lower bounds on the information gap between the prediction task and sensitive information. NoPeek [21] proposes distance correlation minimization as a technique for reducing information leakage and Shredder [14] adds a noise distribution to the intermediate private activations. DeepObfuscator [11] uses adversarial training to prevent reconstruction and attribute leakage attacks. In this work, we focus on selectively pruning intermediate representations to prevent leakage of sensitive information.

Channel pruning is a predominant technique for deep network compression to minimize computational complexity and accelerate inference [4]. While most methods interleave pruning with the training phase [2, 15, 22], there has been recent focus on pruning at inference [7]. [2] gradually prunes channels at fixed intervals during training using a feature relevance score to minimize compute cost. [7] propose dynamic feature boosting and suppression (FBS) to predictively amplify salient convolutional channels and skip unimportant ones at run-time for accelerated inference. In this work, we focus on channel pruning at inference to prevent leakage of sensitive information.

3 Dynamic Channel Pruning Network

In this section we discuss our dynamic channel pruning network for reducing leakage of sensitive attributes from raw input images. As illustrated in Figure 1, the proposed model consists of four key components, namely *client network*, *filter generating network*, *task network*, *adversary network* utilized over two distinct stages.



(a) Performance comparison on the adversary task. The higher it is, more the privacy leakage occurs for the sensitive attributes.

(b) Performance comparison on the target task. The higher it is the better is the performance of the system.

Figure 2: **Evaluation of the proposed method:** we evaluate the performance of our method by baselining it against random pruning and no-pruning procedure. We use fairface as the dataset with two image classification objectives, the adversary task is race prediction (7-class problem) while the target task is gender classification (2-class problem).

Stage 1 - Channel Pruning: First the input sample x goes through the *client network* to generate the intermediate convolutional volume $z \in \mathbb{R}^{h \times w \times c}$. Then, the *filter generating network* takes this intermediate activation z and generate a feature map score $F_\Theta \in \mathbb{R}^c$ for each channel in z. The F_Θ channel pruning filters are weakly discretized using sigmoid with temperature ((to avoid introducing discontinuity) and then thresholded to obtain a binary vector b. This obtained vector b is multiplied channel wise with z to produce a *pruned* feature volume $\hat{z}$ with channels corresponding to the private representations masked out (or deactivated). Note that F_Θ, the feature map score, is conditioned on z (hence x) and is thus generated dynamically on run-time. A hyperparameter *pruning ratio*, is used to define the number of active channels, and helps regulate the privacy-utility trade-off.

Stage 2 - Task Prediction: Next, the pruned activations $\hat{z}$ are consumed by the *task network* and *adversary network* to obtain the corresponding output predictions. Both these predictive networks consist of convolutional and fully-connected layers, in order, to generate their outputs.

Training: The adversary network and task network have access to supervised labels and attempt to minimize their losses. The filter generating network is trained to minimize the loss for the task network and maximize the loss for the adversary network simulating an implicit min-max optimization for these two components. The client network parameters are only optimized with the loss for the target network. We deliberately prohibit gradients flow from adversary to the client network to ensure that the filter generating network generalizes and does not trivially utilize representations learned by the client network. We posit that this facilitates the filter generating network to specialize at pruning by identifying the privacy leaking channels.

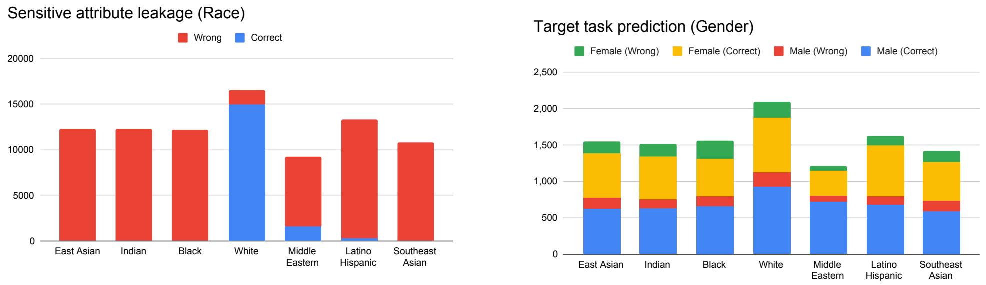
4 Experiments and Results

Experiments are conducted with the Fairface dataset [10] on the task of image classification. The dataset consists of 108,501 face images each tagged with *race*, *age* and *gender* attributes. We specifically select this dataset because of an equitable class balance across different (7) race categories which prevents the trained models from bias towards particular any demographic. We baseline the proposed dynamic channel pruning setup with two configurations:

- *Random Pruning*: Channels are randomly pruned from z conditioned on the *pruning ratio*.
- *No Pruning*: No channels are pruned from z (i.e $\hat{z} = z$)

For scope of this study, experiments use the ResNet-18 [9] architecture with the first 6 convolutional layers as part of the client network. Additional (hyper)parameter details include: Adam optimizer (LR = 1e-2; Batch Size=64) with 4 NVIDIA 1080Ti GPUs.

We experiment with two distinct configuration: a) Prediction Attribute: *Gender* and Sensitive Attribute: *Race* b) Prediction Attribute: *Race* and Sensitive Attribute: *Gender*. The results are



(a) **Performance evaluation for sensitive attribute prediction and studying at risk groups.** Here, the adversary network appears to be predicting majority of the input samples as *White*, except few percentage of samples in the group *Middle Eastern* and *Latino Hispanic* which get correctly identified.

(b) **Evaluating target task performance across different demographics.** We compare performance of the main task (gender classification) segregated by the categories in private task (race classification), so as to study influence of pruned channels for the privacy leakage reduction task.

Figure 3: Experiments dedicated to study ethical considerations of our proposed approach and evaluating whether one particular group of individuals is at more risk than the others.

presented in figure 2. We observe that the proposed dynamic pruning approach achieves significantly better privacy-utility trade-off than the baselines. With race as sensitive attribute (7 class; chance acc: 14%), dynamic pruning results in an accuracy of 20% for the adversary while with random and no-pruning, the adversary network achieves accuracy of 43% and 40% respectively. Further, while achieving superior performance with the adversary network, dynamic pruning also maintains same accuracy as the baselines on task network. We obtain similar trends with gender as sensitive attribute (2 class; chance acc: 50%). Due to limited space, we only include plot for the more challenging (as lower chance accuracy) configuration of race as sensitive attribute.

5 Discussion

Ethical Consideration We additionally focus on granular demographic level privacy-utility trade-offs for the individual race categories in the dataset, *East Asian, Indian, Black, White, Middle Eastern, Latino Hispanic, Southeast Asia*. Results in figure 3a indicate that while the adversary network can correctly classify majority of validation examples in *White* category and small minority of *Middle Eastern* and *Latino Hispanic*, predictions for all other classes are completely erroneous. Furthermore, we extend this same analysis to the target network, being trained on gender prediction in the particular setup. Results in figure 3b show that incorrect predictions are distributed equally across the different race categories.

Privacy-Utility Trade-off In this work we present empirical evidence for the protection of sensitive information, however, for any practically deployable private system, we would require worst case privacy guarantees about the information leakage of data or sensitive attributes. It is important to highlight that the privacy vs utility trade-off depends on what we define as the sensitive task and target task, and how much these two objectives differ. The channel pruning mechanism is based on the hypothesis that each channel roughly disentangles [6] different attributes about the input sample and a subset of them can be pruned.

6 Conclusion

In this work we propose a method for enabling inference on cloud by sharing intermediate activations instead of sharing raw data. To reduce the privacy leakage of the shared data, we perform channel pruning on the client side. The design of the proposed method can go beyond channel pruning for CNNs to a general latent space pruning procedures for feed-forward networks and its other variants used in language models like RNNs and transformers. Future work includes rigorous evaluations of different training configurations and theoretical guarantees to understand privacy and utility trade-off.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2016.

[2] Sai Aparna Aketi, Sourjya Roy, Anand Raghunathan, and Kaushik Roy. Gradual channel pruning while training using feature relevance scores for convolutional neural networks. abs/2002.09958, 2020.

[3] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. *CoRR*, abs/1611.04482, 2016.

[4] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017.

[5] Alexey Dosovitskiy and Thomas Brox. Inverting convolutional networks with convolutional networks. *CoRR*, abs/1506.02753, 2015.

[6] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

[7] Xitong Gao, Yiren Zhao, Łukasz Dudziak, Robert Mullins, and Cheng zhong Xu. Dynamic channel pruning: Feature boosting and suppression. In *International Conference on Learning Representations*, 2019.

[8] Otkrist Gupta and Ramesh Raskar. Distributed learning of deep neural network over multiple agents. *CoRR*, abs/1810.06060, 2018.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[10] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age, 2019.

[11] Ang Li, Jiayi Guo, Huanrui Yang, and Yiran Chen. Deepobfuscator: Adversarial training framework for privacy-preserving image classification, 2019.

[12] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. *CoRR*, abs/1412.0035, 2014.

[13] H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016.

[14] Fatemehsadat Mireshghallah, Mohammadkazem Taram, Prakash Ramrakhyani, Dean M. Tullsen, and Hadi Esmaeilzadeh. Shredder: Learning noise to protect privacy with partial DNN inference on the edge. *CoRR*, abs/1905.11814, 2019.

[15] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[16] Seyed Ali Osia, Ali Taheri, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Hamid R. Rabiee. Deep private-feature extraction, 2018.

[17] Seyed Ali Ossia, Ali Shahin Shamsabadi, Ali Taheri, Kleomenis Katevas, Hamid R. Rabiee, Nicholas D. Lane, and Hamed Haddadi. Privacy-preserving deep inference for rich user data on the cloud. *CoRR*, abs/1710.01727, 2017.

[18] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

[19] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Deep image prior. *CoRR*, abs/1711.10925, 2017.

[20] Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. Split learning for health: Distributed deep learning without sharing raw patient data. *CoRR*, abs/1812.00564, 2018.

[21] Praneeth Vepakomma, Abhishek Singh, Otkrist Gupta, and Ramesh Raskar. Nopeek: Information leakage reduction to share activations in distributed deep learning, 2020.

[22] Zhuangwei Zhuang, Mingkui Tan, Bohan Zhuang, Jing Liu, Yong Guo, Qingyao Wu, Junzhou Huang, and Jinhui Zhu. Discrimination-aware channel pruning for deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 875–886. Curran Associates, Inc., 2018.