INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Differentially Private Adapters for Parameter Efficient Acoustic Modeling

Chun-Wei Ho[1], *Chao-Han Huck Yang*[1], *Sabato Marco Siniscalchi*[1,2,3]

[1]Georgia Institute of Technology, U.S.A
[2]Kore University of Enna, Italy
[3]Norwegian University of Science and Technology, Norway

{chun-wei.ho,huckiyang}@gatech.edu; marco.siniscalchi@ntnu.no

Abstract

In this work, we devise a parameter-efficient solution to bring differential privacy (DP) guarantees into adaptation of a cross-lingual speech classifier. We investigate a new frozen pre-trained adaptation framework for DP-preserving speech modeling without full model fine-tuning. First, we introduce a noisy teacher-student ensemble into a conventional adaptation scheme leveraging a frozen pre-trained acoustic model and attain superior performance than DP-based stochastic gradient descent (DPSGD). Next, we insert residual adapters (RA) between layers of the frozen pre-trained acoustic model. The RAs reduce training cost and time significantly with a negligible performance drop. Evaluated on the open-access Multilingual Spoken Words (MLSW) dataset, our solution reduces the number of trainable parameters by 97.5% using the RAs with only a 4% performance drop with respect to fine-tuning the cross-lingual speech classifier while preserving DP guarantees.

Index Terms: speech classification, differential privacy, domain adaptation, parameter efficient tuning

1. Introduction

With the rapid growth of the computation ability and commercial datasets, more and more personal data are collected, which poses the issue of protecting sensitive data. The United States Census Bureau, for instance, announced a new security standard [1] based on Differential Privacy (DP) [2]. The (ϵ, δ)-DP mechanism allows us to measure the security of algorithms and provides a guarantee based on a privacy budget. However, ensuring differential privacy degrades the system's performance [3] because it restricts access to the data. In addition, training a large model with DP is not only time-consuming but also leads to a more severe drop in performance. [3].

Nonetheless, there are many benefits associated with the use of large-scale datasets and large models. For example, large-scale datasets are fundamental to deploying well-trained deep neural networks (DNNs) [4, 5]; moreover, if the size of the DNN is large enough, it can reach the global minima from any initialization with the gradient descent algorithm [6]. Although the global optimality was only proven in tensor factorization, [6] shows the benefits associated with large connectionist models. Indeed, there exist several large pre-trained models that have been proven vital for different downstream tasks [7, 8, 9, 10, 11, 12, 13, 14] after fine-tuning - in this work, we will use the term fine-tuning and adaptation interchangeably.

Unfortunately, fine-tuning a pre-trained larger model, in addition to being a time-intensive procedure, can also distort the pre-trained features and underperform out-of-distribution [15]. Training large models with differential privacy is even harder because DP-related perturbations are introduced into the training process. Therefore, a feasible solution to estimate and exploit a representation of a large pre-trained model is becoming a pressing issue to be tackled.

This work aims at investigating the benefits of leveraging model adaptation and parameter efficient techniques in the context of differential privacy. In particular, we propose a cross-domain differential private fine-tuning framework [1] leveraging a deep frozen model pre-trained on public source data, and private target data. We consider the case when there is a domain mismatch between source and target domains. In the proposed framework the frozen pre-trained model doesn't guarantee privacy but provides information from non-sensitive source data. We also use additional parameters (weights) to serve as a domain adaptor, which provides information from the target data and introduces DP guarantees. In particular, DP stochastic gradient descent (DPSGD) [16, 17, 18], and Private Aggregation of Teacher Ensembles (PATE) [19, 20] are used to attain DP guarantees.

For DPSGD, we follow what was proposed by Da *et al.* in [21]. Since the experimental evidence demonstrated poor results with DPSGD, we devised a PATE-based solution, which led to a substantial performance improvement. Figure 1 shows the proposed PATE-based solution to perform model adaptation (fine-tuning) while attaining DP-privacy guarantees. The *additional weights* shown in the figure are trained on different disjoint chunks of the sensitive data. Those weights are then inserted into the frozen pre-trained large models using the solutions discussed in [21]. The obtained frozen pre-trained model is aggregated with different weights together based on PATE's algorithm. Finally, the student model queries from the aggregated teacher model using non-sensitive target domain data and learns only from non-sensitive data to preserve privacy. To the best of the authors' knowledge, our work is the first to propose cross-domain DP-based acoustic modeling adaptation. The overall solution does not only guarantee DP, but it also is parameter efficient.

2. Related Works

2.1. Differential Privacy in a Nutshell

The DP mechanism [2] is established to evaluate the security of an algorithm. DP is parameterized by the privacy budget variable ϵ, and δ defined as follows:

Definition 1 An algorithm $\mathcal{A}$ is said to be (ϵ, δ)-DP if for all adjacent datasets D and D', and for any possible event S, the algorithm satisfies:

$$\Pr[\mathcal{A}(D) \in S] \leq e^{\epsilon}\Pr[\mathcal{A}(D') \in S] + \delta \tag{1}$$

[1] GitHub Link: https://github.com/Chun-wei-Ho/Private-Speech-Adapter.

10.21437/Interspeech.2023-551

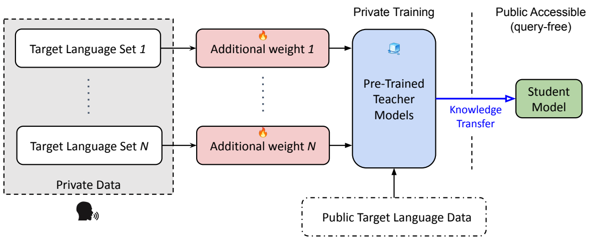


Figure 1: *Proposed private aggregation of teacher ensembles [19] (PATE)-based adapter for parameter efficient fine-tuning on acoustic and speech processing.*

The above equation, in some sense, guarantees that the outcomes of the algorithm with D and D' are indistinguishable.

There are several methods to achieve (ϵ, δ)-DP, and most of them require some DP-oriented perturbation. The perturbation guarantees (ϵ, δ)-DP by making the output of the algorithm, $\mathcal{A}(D)$ and $\mathcal{A}(D')$, indistinguishable. The simplest method to guarantee DP is to introduce the Laplace perturbation to the output of $\mathcal{A}$. It has been shown that we can achieve pure DP ($\delta = 0$) with Laplace perturbation added [22].

Although there exist several ways to estimate the privacy budget ϵ, one of the most convenient methods is Renyi Differential Privacy (RDP) [23], which is based on the Renyi Divergence by (2), which is similar to the Kullback-Leibler Divergence:

$$\mathrm{Div}_\alpha(P\|Q) = \frac{1}{\alpha - 1} E_{X \sim Q} \log \left(\frac{P(x)}{Q(x)} \right)^\alpha \tag{2}$$

The RDP is defined as follows:

Definition 2 An algorithm $\mathcal{A}$ is said to be α, ϵ-RDP if for all adjacent datasets D and D', the algorithm satisfies:

$$\mathrm{Div}_\alpha(f(D) \| f(D')) \leq \epsilon \tag{3}$$

It has been proven in [23] that if any algorithm satisfies α, ϵ-RDP, it's also an $\left(\epsilon + \frac{\log(1/\delta)}{\alpha - 1}, \delta\right)$-DP algorithm. We use RDP to evaluate DP in this study.

2.2. Privacy Preserving in Machine Learning

A common method to preserve privacy is by DP-based perturbations. However, perturbations also degrade the system's performance. Finding a trade-off between performance and privacy has become an important topic worth investigating. Two popular algorithms have been designed to preserve privacy in machine learning. The first is DP stochastic gradient descent (DPSGD) [16, 17, 18], in which the effect of single data is restricted by per-utterance gradient clipping, and the noises are added to satisfy a certain privacy budget ϵ. The second method is PATE [19], which is based on three stages: First, several teacher models are trained on disjoint chunks of sensitive data. Then, the outputs of the teacher models $T_i(x, \theta_i)$ are aggregated using a private aggregation algorithm (4). Finally, the student model is trained on some public data and the output of the teacher models, defined as $T(x, \theta)$ in (4), where $\mathrm{Lap}(\lambda)$ denotes the Laplace perturbation parameterized by λ. PATE models achieve (ϵ, δ)-DP by introducing noises in the aggregation phase and by hiding sensitive data from the student model. The amount of noise is determined by the "smooth sensitivity" [24] of the teacher models, which is also called data-dependent privacy. By reducing the required DP-oriented perturbation while aggregating, PATE has been tested as the state-of-the-art results in different applications, e.g., [25, 26].

$$T(x, \theta) = \mathrm{argmax} \left\{ \left(\sum T_i(x, \theta_i) \right) + \mathrm{Lap}_{i.i.d}(\lambda) \right\} \tag{4}$$

2.3. Parameter Efficiency & Differential Privacy

Training a huge deep model taking into account DP requirements can be troublesome because we have to restrict the information extracted from the data. Furthermore, the perturbation introduces randomness into the learning phase. The amount of perturbation required under the same privacy budget is depended on the model size. The larger the model is, the more perturbation we need to preserve privacy. For example, the perturbation added to the gradients is proportional to the square root of the number of trainable parameters in DPSGD. That in turn leads to a trade-off between the model capacity, and DP guarantees. In many DP setups [19, 27], smaller and simpler model architectures end up providing superior performance. Nonetheless, Da *et al.* [21] proposed to use parameter efficient methods to deal with the noise injection while training large models with DPSGD. In their study, it has been experimentally proven that larger models with parameter efficiency lead to better results when used in combination with DPSGD. We posit that parameter efficiency serves as a conduit between large models and privacy budgets. To this end, we investigate a first attempt to advance parameter-efficient learning with PATE, which has been demonstrated to have wide-ranging applications for performance-driven tasks.

2.4. Parameter Efficient Algorithms

In this study, we mainly focus on two parameter efficient algorithms. Linear Probing (LP) [15] prevents distortions by freezing the entire encoder while training the linear head[2] only. By reusing the pre-trained weights completely, Linear Probing is effective when the source domain and the target domain are similar to each other.

Adapters [28] modifies the feature extractors by inserting some adapting layers without changing the pre-trained weights. More specifically, the relationship between the output of the i^{th} layers $\hat{\mathcal{F}}_\theta^i(x)$ and the output of the $(i-1)^{th}$ layers $\hat{\mathcal{F}}_\theta^{i-1}(x)$ are described in (5), where Θ denotes non-trainable parameters, and θ denotes trainable parameters. $\mathcal{A}_\theta$ denotes some non-linear function parameterized by θ. The hat notation, $\hat{\cdot}$, indicates the functions whose inputs are the model input, x, instead of the output of the previous layer.

$$\begin{aligned} \theta^* &= \arg\min_\theta \left\{ \mathcal{L}_{\text{error}}(\sigma(\hat{\mathcal{F}}_\theta^N(x)), \hat{y}) \right\} \\ \text{where} & \begin{cases} \hat{\mathcal{A}}_\theta^i(x) = & \mathcal{A}_\theta^i(\hat{\mathcal{F}}_\theta^{i-1}(x)) \\ \hat{\mathcal{F}}_\theta^i(x) = & \underbrace{\mathcal{F}_\Theta^i(\hat{\mathcal{F}}_\theta^{i-1}(x))}_{\text{original encoder (frozen)}} + \underbrace{\hat{\mathcal{A}}_\theta^i(x)}_{\text{Adapter output}} \end{cases} \end{aligned} \tag{5}$$

DNN Residule Adapter ($\mathrm{RA}_{\mathrm{DNN}}$) [29], a common adapter uses a simple up-projector and a simple down-projector along with a residual path to define the non-linear function $\mathcal{A}_\theta$, which modifies the input feature, $\hat{\mathcal{F}}_\theta^{i-1}(x)$, by a limited matrix rank. It has been experimentally proven that $\mathrm{RA}_{\mathrm{DNN}}$ can attain comparable performance results to those obtained through a fine-tuning of the whole model parameters but using only up to 2 % of parameters [30].

[2] The last linear layer is referred to as "head"

3. Proposed DP based Parameter Efficient Adaptation for Acoustic Modeling

In this study, two of the most popular privacy-preserving algorithms, DPSGD and PATE were investigated. For DPSGD, we used the same setup in [21], where only the RA_{DNN}s are updated during training. Figure 1 shows instead the proposed PATE-based solution, where N different additional weights are trained on the different disjoint chunks from the sensitive dataset. The weights are then inserted into the global teacher model and are aggregated together using the private aggregation algorithm proposed in [19]. The student model, on the other hand, learns from the public data queried from the private teacher models. Therefore, the student can learn from private data without direct access to it. As explained in Section 2.2, the amount of required DP-oriented perturbation is determined by the sensitivity of the teacher models. Therefore, by applying data-dependant privacy and domain adaptation, we were able to successfully reduce the amount of DP-oriented perturbation required to preserve privacy.

3.1. DNN Residual Adapters Connection

As discussed in Section 2.4, RA_{DNN} is one of the common parameter-efficient adapters. In this study, we also investigated different non-linear functions, $\hat{\mathcal{A}}_\theta(x)$. Inspired by [31, 32], we try to connect the RA_{DNN}s using some skip connections. Instead of just performing neighboring connections, we tried to connect the RA_{DNN}s in three different ways and investigate their effects. The three connection ways are summarized in (6). The connections are inspired by Unet [33] and DenseNet [34].

$$
\begin{aligned}
&\text{Neighboring:} \quad \hat{\mathcal{A}}^i_\theta(x) = \mathcal{A}^i_\theta(\hat{\mathcal{F}}^{i-1}_\theta(x) + \hat{\mathcal{A}}^{i-1}_\theta(x)) \\
&\text{Unet-alike [33]:} \quad \hat{\mathcal{A}}^i_\theta(x) = \mathcal{A}^i_\theta(\hat{\mathcal{F}}^{i-1}_\theta(x) + \hat{\mathcal{A}}^{N-i}_\theta(x)) \; \forall i > \frac{N}{2} \\
&\text{DenseNet-alike [34]:} \quad \hat{\mathcal{A}}^i_\theta(x) = \mathcal{A}^i_\theta(\hat{\mathcal{F}}^{i-1}_\theta(x) + \sum_{k=1}^{i-1} \hat{\mathcal{A}}^k_\theta(x))
\end{aligned}
\tag{6}
$$

As defined in (6), the neighboring connections connect the output of the previous layer. In the Unet-alike connection, the last i layers are connected to the first i layers. And in the DenseNet-alike connection, every layer is connected to every preceding layer.

3.2. Evaluation of Utility

We leveraged Eric Hulburd's work [35] to assess the quality of the proposed approach and used the utility defined in (7) that takes both parameter efficiency and performance:

$$
\text{Utility} = \frac{\text{Accuracy} - 50}{\log(\text{Number of trainable parameters})} \tag{7}
$$

4. Experiments & Results

4.1. Experimental Setup

We assessed our framework on a keyword classification task. Specifically, we used the English Google Speech Command V2 (EGSP-V2) [36] as source domain, and the Multilingual Spoken Words [37] as the target domain. We took into account only four languages, namely English, German, French, and Russian, and generated smaller subsets from them, referred to as *MLSW-mini* [3], to simulate low-resource conditions. *MLSW-mini* configuration is shown in Table 1. And the EGSP-V2 was used to pre-train the deep classifier. Then, we adapted the model to *MLSW-mini* with DP. For DPSGD, we used *MLSW-mini-train* and half of *MLSW-mini-test* to train the model. For PATE, we trained the teacher models on *MLSW-mini-train.* Then we trained the student model on half of the *MLSW-mini-test.* The remaining data in *MLSW-mini-test* was used for evaluation. The proposed setup follows the standard PATE setup [19]. The privacy budget ϵ is 8.0 [4] for French, German, and English, and 11.6 for Russian.

Table 1: *MLSW-mini dataset. "# Words" indicates the number of unique words in the language. The sample rate of the waveforms is 16 kHz. Each waveform is roughly 1 second long.*

Language	# Words	# Samples/word	Total Train Audio Time
en (Germanic)	18	4501-4927	23 hours 34 mins
de (Germanic)	15	4011-4910	18 hours 14 mins
fr (Romance)	13	4081-4988	16 hours 01 mins
ru (Slavic)	23	1002-4758	11 hours 00 mins

The deep architecture used as a pre-trained model is the Keyword Transformer (KWT) [38]. KWT first performs a time-distributed linear project of the mel-spectrogram; it then concatenates the features with token embeddings. Next, the concatenated features are fed into 12 layers of transformer blocks with dimension 192 and classified using a linear head. The setups are similar to [38] with the only difference being that 12 trainable RA_{DNN}s (with different dimensions) were inserted between the transformer blocks in the fine-tuning phase.

The Mel-spectrogram generated with a 30 ms analysis window, a 10 ms frame shift, and 40-points DFT is the input feature used in both pre-training, and fine-tuning. For optimization, we used AdamW except for DPSGD. The number of epochs was set to 200. All the other setups are the same as those in [38].

4.2. Cross-lingual Adaptation Results

In this section, we investigate the effect of domain adaption with cross-lingual data and compare the two introduced DP algorithms, DPSGD and PATE, with and without RA_{DNN}s.

In Table 2, the baseline method, i.e., adapting the whole KWT network parameters without DP guarantees, attains a classification accuracy equal to 96.49 with a utility of 3.0 on the France language. By comparing the results with or without DP in Table 2, we can see that both utility and accuracy drop when DP constraints are imposed. In particular, the accuracy drops from 96.49 to 53.40 when DPSGD, the fourth row, is used, and the utility drops from 3.0 to 0.22. PATE, in the fifth row, instead can limit the drop in accuracy and utility.

Furthermore, by comparing the results of training from scratch (fs) and fine-tuning (FT), we conclude that domain adaptation is required to successfully train a model with DP. But differ from what was reported on language modeling in [21], DPSGD is not effective in cross-lingual acoustic adaptation but PATE is. We argue the difference is mainly because the domain mismatch is larger in cross-lingual tasks, and the mechanism of data dependant privacy in PATE reduces the amount of perturbation needed to be added under the same level of privacy budget. We also evaluate all the selected languages listed in Table 1, reporting an overall average results in the last two rows in

[3] The list of train/test split is reported on GitHub.

[4] We follow a common privacy budget (ϵ=8) based on [21] and Apple's official document in `https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf`

Table 2: *Comparison between DPSGD and PATE with and without RA$_{DNN}$s on MLSW-mini. The All language results are the weighted average accuracy based on the number of utterances in the four selected languages.*

lang	Method	DP	# Train Para.	Utility	Acc. (%)
fr	from Scratch (fS)	✗	5.4 M (100 %)	2.88	94.58
	fS w/ PATE	✓	5.4 M (100 %)	2.66	91.23
en → fr	Fine-tune (FT)	✗	5.4 M (100 %)	3.00	96.49
	FT w/ DPSGD	✓	5.4 M (100 %)	0.22	53.40
	FT w/ PATE	✓	5.4 M (100 %)	2.72	92.10
	LP w/ PATE	✓	21.7 K (0.4 %)	1.13	61.13
	RA$_{DNN}$ w/ DPSGD	✓	0.9 M (14.6 %)	0.77	60.69
	RA$_{DNN}$ w/ PATE	✓	0.9 M (14.6 %)	**3.05**	91.82
all	fS	✗	5.4 M (100 %)	2.71	92.03
	fS w/ PATE	✓	5.4 M (100 %)	2.32	85.97
en → all	FT	✗	5.4 M (100 %)	2.99	96.38
	FT w/ PATE	✓	5.4 M (100 %)	2.49	88.51
	RA$_{DNN}$ w/ PATE	✓	0.9 M (17 %)	**2.75**	87.72

Table 3: *Results with RA$_{DNN}$ for PATE with different RA$_{DNN}$ dimension and a privacy budget* $\epsilon = 7.96$ *on MLSW-mini French. RA$_{DNN\text{-}d}$ means the down-projection dimension is d.*

Method	DP	# Train Para.	Utility	Acc. (%)
FT	✗	5.4 M (100%)	3.00	96.49
FT w/ PATE	✓	5.4 M (100%)	2.72	92.10
LP	✓	21.7 K (0.4%)	1.13	61.13
RA$_{DNN\text{-}24}$	✓	0.1 M (2.5%)	**3.22**	88.08
RA$_{DNN\text{-}288}$	✓	1.4 M (20.2%)	2.98	**92.07**

Table 2. The results indicate that our method works not only on French but also in multi-lingual scenario.

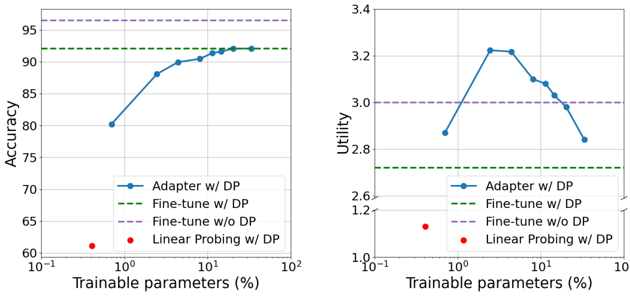
4.3. Residual Adapter Size Effect on Fine-tuning

In this section, the effects of RA$_{DNN}$s are discussed. The *MLSW-mini* French in Table 1 is used for our experiment. We performed the experiments with a privacy budget $\epsilon = 7.96$ using a PATE [19] based KWT [38] model. We also used use RA$_{DNN\text{-d}}$ to denote that the down-projection dimension of the RA$_{DNN}$s is *d*. We tried several *d* values ranging from 3 to 288, where 288 is twice the dimension of the original feature dimension, and 3 is instead 64 times smaller than the original feature dimension. As summarized in Table 3, RA$_{DNN\text{-}24}$ attains the best utility, with an 88.08 % accuracy training 2.46 % of parameters only. In addition, by appropriately choosing the size of RA$_{DNN}$s, RA$_{DNN\text{-}288}$ provides a result that is comparable with that of the fully fine-tuned model.

The effect of trainable parameters is also investigated in Figure 2. First of all, as the number of trainable parameters increases, the model accuracy increases. However, it saturates at the fine-tuned accuracy when the number of parameters exceeds 20% of the total model parameters. The latter means that we only have to train 20% of the parameters to reach the best performance, and increasing the number of trainable parameters does not help. In addition, the best utility occurs when adapting 2.46% of parameters. Reducing it does not lead to any beneficial effect, and the accuracy begins to degrade rapidly. Increasing the RA$_{DNN}$ size improves the overall accuracy, but the utility drops because the number of trainable parameters increases accordingly.

4.4. Different Connections of Residual Adapters

We now investigate into the different RA$_{DNN}$ connections described in Section 3.1. As shown in Table 4, connecting the RA$_{DNN}$s in our task isn't necessarily helpful. We believe the reason is that the additional information from the other RA$_{DNN}$s is too noisy for a few-shot domain adaptation. We can validate the hypothesis from the fact that the DenseNet-alike connections provide the worst performance albeit it's more complicated than the other structures. And the results, same as our other experiments, lead to a conclusion that the simpler, the more promising.



(a) *Accuracy* (b) *Utility*

Figure 2: *Accuracy and utility of PATE-RA$_{DNN}$ architecture with different RA$_{DNN}$ sizes. (a) The model performance converges to the fine-tuning result when 20 % of the parameters are adapted. (b) Our method achieves the best utility when 2.46 % of parameters are adapted.*

Table 4: *Experiments of PATE with different connections from EGSP-V2 to MLSW-mini French with* $\epsilon = 8.0$.

Model structure	Connection type	Acc. (%)
RA$_{DNN\text{-}24}$	No connection	**88.08**
	Neighboring [31]	86.91
	Unet-alike	87.61
	DenseNet-alike	86.72
RA$_{DNN\text{-}288}$	No connection	**92.07**
	Neighboring [31]	91.49
	Unet-alike	91.77
	DenseNet-alike	91.13

5. Conclusion

In this work, we tackled the problem of preserving privacy in a cross-lingual speech classification task. First, we tried to port what done on language modeling by [21] using DPSGD, but we observed a significant performance drop using their method. Thus, we proposed a novel PATE-based solution, which, differently from DPSGD, led to a small drop in performance while still preserving DP guarantees.

Furthermore, to reduce the computational burden while fine-tuning with DP, we tested LP and RA$_{DNN}$. LP was not effective; whereas, RA$_{DNN}$ allows a reduction of 97.5% of the parameters to be adapted while keeping a comparable performance of the PATE model. We also performed an ablation study to verify skip connection strategies on RA$_{DNN}$. Although skip-connection does not give any performance improvement, the exploring of different parameter-efficient architectures leveraging PATE is useful for future studies.

Acknowledgments The authors would like to express their gratitude to Prof. Chin-Hui Lee from Georgia Tech for providing helpful insights and suggestions.

6. References

[1] U. Bureau, "Disclosure avoidance for the 2020 census: An introduction," 2021.

[2] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings 5*. Springer, 2008, pp. 1–19.

[3] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," *Advances in neural information processing systems*, vol. 32, 2019.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] S. Jean, S. Lauly, O. Firat, and K. Cho, "Does neural machine translation benefit from larger context?" *arXiv preprint arXiv:1704.05135*, 2017.

[6] B. D. Haeffele and R. Vidal, "Global optimality in tensor factorization, deep learning, and beyond," *arXiv preprint arXiv:1506.07540*, 2015.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[9] A. Radford, J. W. Kim *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[11] B. Li, D. Hwang, Z. Huo, J. Bai, G. Prakash, T. N. Sainath, K. C. Sim, Y. Zhang, W. Han, T. Strohman *et al.*, "Efficient domain adaptation for speech foundation models," in *Proc. of ICASSP*. IEEE, 2023, pp. 1–5.

[12] Y.-N. Hung, C.-H. H. Yang, P.-Y. Chen, and A. Lerch, "Low-resource music genre classification with cross-modal neural model reprogramming," in *Proc. of ICASSP*. IEEE, 2023.

[13] K.-W. Chang, Y.-K. Wang, H. Shen, I.-t. Kang, W.-C. Tseng, S.-W. Li, and H.-y. Lee, "Speechprompt v2: Prompt tuning for speech classification tasks," *arXiv preprint arXiv:2303.00733*, 2023.

[14] H. Yen, P.-J. Ku, C.-H. H. Yang, H. Hu, S. M. Siniscalchi, P.-Y. Chen, and Y. Tsao, "Neural model reprogramming with similarity based mapping for low-resource spoken command classification," *arXiv preprint arXiv:2110.03894*, 2021.

[15] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," *arXiv preprint arXiv:2202.10054*, 2022.

[16] S. Song, K. Chaudhuri, and A. D. Sarwate, "Stochastic gradient descent with differentially private updates," in *2013 IEEE global conference on signal and information processing*. IEEE, 2013, pp. 245–248.

[17] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *2014 IEEE 55th annual symposium on foundations of computer science*. IEEE, 2014, pp. 464–473.

[18] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[19] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *arXiv preprint arXiv:1610.05755*, 2016.

[20] C.-H. H. Yang, S. M. Siniscalchi, and C.-H. Lee, "Pate-aae: Incorporating adversarial autoencoder into private aggregation of teacher ensembles for spoken command classification," *Proc. of Interspeech*, 2021.

[21] D. Yu, S. Naik *et al.*, "Differentially private fine-tuning of language models," *Proc. of ICLR*, 2022.

[22] R. Sarathy and K. Muralidhar, "Evaluating laplace noise addition to satisfy differential privacy for numeric data." *Trans. Data Priv.*, vol. 4, no. 1, pp. 1–17, 2011.

[23] R. K. Moore and L. Skidmore, "On the use/misuse of the term 'Phoneme'," in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 2340–2344.

[24] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, 2007, pp. 75–84.

[25] J. Jordon, J. Yoon, and M. Van Der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees," in *International conference on learning representations*, 2019.

[26] A. Aslan, T. Matschak, M. Greve, S. Trang, and L. Kolbe, "At what price? exploring the potential and challenges of differentially private machine learning for healthcare," 2023.

[27] F. Tramer and D. Boneh, "Differentially private learning needs better features (or much more data)," *arXiv preprint arXiv:2011.11660*, 2020.

[28] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," *Advances in neural information processing systems*, vol. 30, 2017.

[29] K. Tomanek, V. Zayats, D. Padfield, K. Vaillancourt, and F. Biadsy, "Residual adapters for parameter-efficient asr adaptation to atypical and accented speech," *Proc. of EMNLP*, 2021.

[30] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[31] C.-H. H. Yang, B. Li, Y. Zhang, N. Chen, R. Prabhavalkar, T. N. Sainath, and T. Strohman, "From english to more languages: Parameter-efficient model reprogramming for cross-lingual speech recognition," *Proc. of ICASSP*, 2023.

[32] L. Yang, A. S. Rakin, and D. Fan, "Rep-net: Efficient on-device learning via feature reprogramming," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 277–12 286.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[35] E. Hulburd, "Exploring bert parameter efficiency on the stanford question answering dataset v2. 0," *arXiv preprint arXiv:2002.10670*, 2020.

[36] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[37] M. Mazumder, S. Chitlangia, C. Banbury, Y. Kang, J. M. Ciro, K. Achorn, D. Galvez, M. Sabini, P. Mattson, D. Kanter *et al.*, "Multilingual spoken words corpus," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[38] A. Berg, M. O'Connor, and M. T. Cruz, "Keyword transformer: A self-attention model for keyword spotting," *arXiv preprint arXiv:2104.00769*, 2021.