Personalized Federated Learning for Text Classification with Gradient-Free Prompt Tuning

Anonymous ACL submission

Abstract

In this paper, we study personalized federated learning for text classification with Pretrained Language Models (PLMs). We identify two challenges in efficiently leveraging PLMs for personalized federated learning: 1) *Communication.* PLMs are usually large in size, *e.g.*, with hundreds of millions of parameters, inducing huge communication cost in a federated setting. 2) *Local Training.* Training with PLMs generally requires back-propagation, during which memory consumption can be several times that of the forward-propagation. This may not be affordable when the PLMs are trained locally on the clients, since the clients may be resource constrained, *e.g.*, mobile devices with limited access to memory resources. Additionally, the PLMs can be provided as concealed APIs, for which the back-propagation operations may not be available. For the first challenge, we adopt prompt tuning for PLMs that only train with the prompt parameters, while the pretrained parameters are frozen. We further propose a compression method for the learned prompts to reduce communication cost. For the second challenge, we propose a gradient-free approach based on discrete local search with natural language tokens, circumventing gradient computation with back-propagation, while also reducing the communication cost. Experiments on multiple datasets demonstrates the effectiveness of our method.

1 Introduction

Personalized federated learning (Fallah et al., 2020; Chen et al., 2018; Shamsian et al., 2021) involves collaborative training with non-shareable private data from multiple clients. For each client, we aim to train a personalized model that fits to its own local data, leveraging knowledge from other clients. Personalized federated learning has been attracting increasing attention in the federated learning community due to its ability to account for data heterogeneity across clients (Li et al., 2021). On the other hand, the advent of Pretrained Language Models (PLMs) (Liu et al., 2019; Kenton and Toutanova, 2019) has yielded remarkable performance for tasks involving natural language processing, *e.g.*, text classification. However, such PLMs are usually large in size, *e.g.*, with hundreds of millions of parameters. There has been limited works investigating how to efficiently train with such large PLMs in the federated learning scenarios (Guo et al., 2022; Zhao et al., 2022). In this paper, we investigate on efficient training with PLMs in personalized federated learning for the task of text classification.

One challenge of training PLMs in a federated learning scenario is how to reduce communication cost. Federated learning generally requires communicating updated trainable model parameters between a central server and all the clients (McMahan et al., 2017; Li et al., 2020). When training with PLMs, their sheer size may introduce huge communication cost between the server and clients, thus reducing the training efficiency. To solve this problem, recent works propose to leverage prompt tuning (Guo et al., 2022; Zhao et al., 2022). Specifically, prompt tuning learns with a sequence of trainable prompt embeddings inserted into the input layer of the PLMs. By only training and communicating the prompt embeddings and freeze the pretrained parameters of the PLMs, the communication cost is largely reduced compared with training all the parameters of the PLMs. However, in these works prompt tuning is not realistic for federated learning. The main reason is that the local training, *i.e.*, when the PLMs are trained locally on each client, requires back-propagating through the PLMs in order to calculate the gradient of the prompt embeddings. The memory consumption of back-propagating is several times higher (depending on implementation) than that of forward-propagation[1] (Baydin et al., 2022; Belouze, 2022).

[1] This is because back-propagation requires saving the in-

Such memory consumption is proportional to the size of the PLM, *e.g.*, with hundreds of millions of parameters. Therefore, back-propagating with the PLMs can be extremely memory consuming. Unfortunately, the clients in federated learning usually have limited access to the resources (Rabbani et al., 2021; Deng, 2019), *e.g.*, edge devices with limited memory. As a result, the memory footprint during the local training with back-propagation can exceed the memory capacity of the client devices, making the training infeasible. Further, the PLMs may be provided as concealed APIs, for which the back-propagation operation may not be available (Sun et al., 2022b).

To address those issues, we propose a framework for personalized federated learning that tunes the prompts in the input layer with gradient-free training that only requires forward-propagation. The propose framework consumes less memory making it suitable for federated learning and does not have the limitation from black-box API setting. Our framework follows the conventional two-step training stages; *1) Joint training* - A global model is trained with data from all the clients with federated learning, *2) Post tuning* - Local training that fine tunes the global model from joint training with the local data of each client to learn its personalized model (Fallah et al., 2020; Chen et al., 2018), but in a novel way to enable gradient-free approaches. Specifically, for the joint training, we propose a gradient-free prompt tuning mechanism for the local training of federated learning, based on discrete local search with natural language tokens of the PLMs. By keeping the prompts from local training to be natural language tokens, each client only needs to upload the token indices of the learned prompts to the server, thus significantly reducing the upload communication cost relative to uploading the learned prompt embeddings. We further propose a compression mechanism that reduces the download communication cost. For post-tuning, we adopt black-box tuning (Sun et al., 2022b), which is also gradient-free without back-propagation. Our contributions are as follows:

- We propose a gradient-free personalized federated learning framework for text classification with PLMs. To the best of our knowledge, we are the first to consider gradient-free training in federated learning with PLMs.
- We evaluate the proposed approach on various datasets for text classification. Results show that our approach can achieve superior results for personalized federated learning, along with substantially less communication cost and memory consumption compared with the baselines.

termediate results of a computational graph, while the forward-propagation does not.

2 Related Work

Federated Learning with PLMs: As mentioned above, the sheer size of the PLMs (Liu et al., 2019; Kenton and Toutanova, 2019) poses challenges when applying them to federated learning due to both high communication cost and large memory footprint during local training. Previous works studying PLMs under the federated learning setting only consider the training efficiency in terms of the communication cost, but rarely account for memory footprint. For instance, Lit et al. (2022) propose to reduce the communication cost by only communicating the lower layers of the PLMs between server and clients. Inspired by the superior performance and efficiency of prompt tuning (Lester et al., 2021; Liu et al., 2022) over tuning pretrained parameters, (Guo et al., 2022; Zhao et al., 2022) propose to further reduce the communication cost via only training and communicating the continuous prompt embeddings, trained with gradient descent. The drawback of these works is that they all require gradient computing with back-propagation, which ignores the huge memory consumption caused by back-propagation through the PLMs. As mentioned before, this can be problematic for clients with constrained computation resources, *e.g.*, edge devices with limited memory capacity. Additionally, the PLMs can be provided as concealed APIs (Sun et al., 2022b), for which the back-propagation operation may not be available. (Wang et al., 2020; Dong et al., 2022; Gao et al., 2019) study metric learning and contrastive learning for text representations, which are inspiring for federated learning with text data. However, these works are not targeting federated learning.

Gradient-Free Training with PLMs: Sun et al. (2022b) assumes the PLMs are concealed in black-box APIs and propose to train the input prompt embeddings of the PLMs with CMA-ES (Hansen and Ostermeier, 2001), a gradient-free method that only requires forward-propagation. This setting is termed Language-Model-as-a-Service (LMaaS), where the client data is transferred to an external server with the API of PLMs. This violates the privacy-preserving principle of federated learning. Sun et al. (2022a) further considers gradient-free training with prompts inserted into the intermediate layer of the PLMs, which contradicts our assumption about black-box APIs. Deng et al. (2022); Diao et al. (2022) model the prompts of the inputs layer of PLMs with a prompt generator, and trains them with reinforcement learning. In this way, the back-propagation is not with the PLMs in the API but with the prompt generator. This may not be suitable for federated learning, since adding and back-propagating with prompt generators (*e.g.*, implemented with another PLM) introduce additional memory consumption for clients during local training. Hou et al. (2022); Prasad et al. (2022) also study gradient-free training of PLMs, but it is unclear how to apply their approach for federated learning. Specifically, Hou et al. (2022) adopts boosting with prompts, requiring ten times the computation for model inference compared to without boosting, thus is not compatible with clients equipped with constrained computation resources. Importantly, none of the above works are studying federated learning.

3 General Setup

Let M be the number of clients in federate learning, and $\{\mathcal{D}_1, \dots, \mathcal{D}_M\}$ be the local datasets for all the clients. In personalized federated learning, these datasets are from different domains or tasks. We have $\mathcal{D}_i = \{\boldsymbol{X}_n, \boldsymbol{Y}_n\}_{n=1}^N$, for $i = 1, \dots, M$ with totally N training samples, where $\boldsymbol{X}_n$ is the n^{th} text sequence and $\boldsymbol{Y}_n$ is its label for text classification. Let $\boldsymbol{f}_i(\cdot)$ be the model for client i, with $\boldsymbol{f}_i(\boldsymbol{X}_n)$ being the predicted probability distribution for $\boldsymbol{X}_n$ over all possible labels in client i. The model $\boldsymbol{f}_i$ is implemented as prompt tuning. Specifically, let $\boldsymbol{H}$ be the pretrained encoder of the PLM and $\boldsymbol{p}_i \in \mathbb{R}^{T \times D}$ represent a sequence of T prompt token embeddings. In experiments, we follow (Sun et al., 2022b) that set $T = 50$. D is the dimension of the pretrained token embeddings. $\boldsymbol{f}_i(\boldsymbol{X}_n)$ can be written as,

$$Temp = [\boldsymbol{p}_i; \boldsymbol{e}(\boldsymbol{X}_n); \boldsymbol{e}(It\ is\ [MASK])] \quad (1)$$

$$\boldsymbol{f}_i(\boldsymbol{X}_n) = \mathrm{softmax}(\boldsymbol{H}(Temp) \cdot \boldsymbol{V}_l^T), \quad (2)$$

where $[;]$ denotes row concatenation, $\boldsymbol{p}_i$ is the learnable prompt, $\boldsymbol{e}(\cdot)$ is the embedding layer of the PLM that convert each token in $\boldsymbol{X}_n$ into a token embedding. $\boldsymbol{H}$, and $\boldsymbol{e}$ are frozen during prompt tuning. (1) defines the template for the text classification input, which contains a *[MASK]* token. The output from $\boldsymbol{H}$ on the position of *[MASK]* is compared via inner product with the *verbalizer* $\boldsymbol{V}_l$, which contains embeddings of words that are representative of each label. For instance, we can have $\boldsymbol{V}_l = \boldsymbol{e}([good, bad])$ for sentiment classification.

We see that the only trainable parameter in $\boldsymbol{f}_i(\cdot)$ is the prompt $\boldsymbol{p}_i$. The training loss for client i is,

$$\mathcal{L}(\boldsymbol{p}_i; \mathcal{D}_i) = \frac{1}{N} \sum_{n=1}^{N} \text{cross\_entropy}(\boldsymbol{f}_i(\boldsymbol{X}_n), \boldsymbol{Y}_n), \quad (3)$$

When training with personalized federated learning for text classification, the general objective is to find $\{\boldsymbol{p}_i\}_{i=1}^N$ that minimizes,

$$\frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(\boldsymbol{p}_i; \mathcal{D}_i), \quad (4)$$

while keeping $\{\mathcal{D}_i\}_{i=1}^M$ locally for each client. We follow Sun et al. (2022b) that assumes the pretrained encoder $\boldsymbol{H}$ is concealed in a black-box API whose parameters are not accessible (no parameter leakage) and cannot be backpropagated.

4 Our Framework

As mentioned in Section 1, our framework for personalized federated learning follows (Fallah et al., 2020; Chen et al., 2018), subjecting to a global prompt that is first learned with federated learning (*Joint Training*) and then fine tuned separately with the local data of each client to encourage personalization (*Post Tuning*) . One difference between our approach and (Fallah et al., 2020; Chen et al., 2018) is that our approach focuses on efficient training with gradient-free methods, *i.e.*, without gradient computation using back-propagation. Alternatively, Fallah et al. (2020); Chen et al. (2018) are gradient-based and requires computing second-ordered gradient during joint training with meta-learning, *i.e.*, via MAML (Finn et al., 2017). It remains an open question of how to efficiently estimate the second-ordered gradient without back-propagation, which is out of the scope of our work.

Additionally, compared with previous works of prompt tuning (Li and Liang, 2021; Sun et al., 2022b), our approach improves the training efficiency of federated learning. Specifically, we propose a discrete local search mechanism (see Section 4.1.2) that reduces the upload communication cost in federated learning, while considering personalized federated learning. We also propose a compression method (see Section 4.1.3) that reduces the download communication cost.

4.1 Joint Training

4.1.1 Federated Learning

Below we introduce the general procedure of federated learning with joint training. The goal is to train a global model $\boldsymbol{f}(\cdot)$ with prompt $\boldsymbol{p} \in \mathbb{R}^{T\times D}$ with data from all the clients. Unlike (4), $\boldsymbol{p}$ is expected to minimize the following objective,

$$\frac{1}{M}\sum_{i=1}^{M}\mathcal{L}(\boldsymbol{p};\mathcal{D}_i), \tag{5}$$

which can be optimized with federated learning (McMahan et al., 2017), as shown in Algorithm 1. The federated learning algorithm generally consists of three steps: *1)* Client update; *2)* Aggregation; and *3)* Download. Our proposed gradient-free client update is introduced in Section 4.1.2. For each round of federated learning, given the prompts $\{\boldsymbol{p}_i\}_{i=1}^{M}$ from the client update, the server will aggregate these prompts to generate the global prompt $\boldsymbol{p}$ for the current round, *i.e.*,

$$\boldsymbol{p} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{p}_i, \tag{6}$$

where we adopt FedAvg (McMahan et al., 2017) and assume uniform weighting for each client. The resulting $\boldsymbol{p}$ should be downloaded to each client for the next round of federated learning. Section 4.1.3 proposes a compression method that represents $\boldsymbol{p}$ with reduced memory footprint before download. Note that we assume the API of the PLM has been downloaded to each client before the start of federated learning, so that we only need to communicate the prompts during federated learning. We claim that downloading the API to clients is a practical assumption. This is because it avoids the necessity of uploading client data to an external server (with API) for model inference, compared with the recent Language-Model-as-a-Service (Sun et al., 2022b) where the API is only store on an online server. This is especially important for federated learning where the privacy is of prime concern.

4.1.2 Gradient-Free Client Update

In updating each client i, its prompt $\boldsymbol{p}_i$ is firstly initialized with the global prompt $\boldsymbol{p}$ (or $\boldsymbol{p}'$ in Section 4.1.3) from the previous round of federated

Algorithm 1 Algorithm for Joint Training.

Input: Datasets $\{\mathcal{D}_i\}_{i=1}^{M}$, the PLM (API and its pretrained embedding matrix $\boldsymbol{e}(\mathcal{V})$).
Output: The resulting prompt $\boldsymbol{p}'$.

```
Initialize p with natural token embeddings.
p = p' = p'_{-1}
Download the PLM API and p'_{-1} to each client.
% General procedures for federated learning.
for r = 1, ··· , n_round do
    % Update p_i with each client.
    for i = 1, ··· , M do
        % Please refer to Alg. 3 and Section 4.1.2.
        p_i = Client_Update(p', D_i)
    end for
    % Aggregation.
    Aggregate {p_i}_{i=1}^M with (6), generating p.
    % Please refer to Alg. 2 and Section 4.1.3.
    p' = Compress_Download(p, e(V))
    p'_{-1} = p'
end for
```

learning, then fine tuned on the local dataset $\mathcal{D}_i$. As mentioned before, gradient-based fine tuning of $\boldsymbol{p}$ with back-propagation can be extremely memory consuming with PLMs. Additionally, the back-propagation operation may not be available for PLMs concealed behind APIs. So motivated, we study gradient-free client update of the prompt $\boldsymbol{p}$, which does not need gradient computation with back-propagation and is compatible with the APIs. Specifically, we propose an update mechanism based on discrete local search with natural language tokens. Let $\mathcal{V}$ be the vocabulary of the PLM and superscript t denote the t^{th} row of a matrix. For each iteration update, given a randomly sampled position of the prompts t, $t \in [1, T]$, and a set of candidate tokens $\boldsymbol{C} \subset \mathcal{V}$, we update $\boldsymbol{p}_i^t$ via,

$$\boldsymbol{p}_i^t = \underset{\boldsymbol{w}\in\{\boldsymbol{p}_i^t\}\cup\{\boldsymbol{e}(c)|c\in\boldsymbol{C}\}}{\operatorname{argmin}} \mathcal{L}(\operatorname{rep}(\boldsymbol{p}_i,\boldsymbol{w},t),\mathcal{D}_i), \tag{7}$$

Note that $\boldsymbol{p}_i^t$ on the left side is the updated prompt of the next iteration, while the one on the right is that of the previous iteration. Further, $\operatorname{rep}(\boldsymbol{p}_i,\boldsymbol{w},t)$ denotes replacing the t^{th} row of $\boldsymbol{p}_i$ with $\boldsymbol{w}$. We randomly choose one position t for each update iteration. The candidate set $\boldsymbol{C}$ is selected with,

$$\boldsymbol{C} = \underset{\boldsymbol{C}\subset\mathcal{V},|\boldsymbol{C}|=K}{\operatorname{argmin}} \sum_{c\in\boldsymbol{C}}\cos(\boldsymbol{e}(c),\boldsymbol{p}_i^t), \tag{8}$$

where $\cos(\cdot)$ is the cosine distance. We only select K candidate tokens in $\boldsymbol{C}$ with the most similar semantics as $\boldsymbol{p}_i^t$ (low cosine distance), which avoids large change of $\boldsymbol{p}_i^t$ in a single iteration. K is the number of local search for each step that controls the training efficiency and is discussed in Section 5. The general procedures are shown in Algorithm 3.

Such a simple update mechanism has two benefits. Firstly, since $\boldsymbol{w}$ on the right side of (7) can take the value of $\boldsymbol{p}_i^t$, the value of $\mathcal{L}(\boldsymbol{p}_i, \mathcal{D}_i)$ should be non-increasing during client update. Secondly, by constraining the candidate embeddings to be from the natural language tokens, *i.e.*, $\boldsymbol{C} \subset \mathcal{V}$, the updated positions of $\boldsymbol{p}_i$ can be saved by only keeping its token index. This significantly reduces the communication cost when uploading prompts to the server, compared with previous works of continuous prompt tuning Guo et al. (2022); Zhao et al. (2022) that upload all the prompt parameters. For instance, the vocabulary size of the Roberta-Large (Liu et al., 2019) model is 50,264 with $D = 1024$, which implies that each token index can be encoded with 16 bits. For positions of $\boldsymbol{p}_i$ that are not modified during client update, we can indicate it with a special index using a 16-bit integer, *e.g.*, 50,265 (not natural token indices). Thus, we only need to upload 16 Bits for each position of p_i. Comparatively, uploading the whole prompt vector to the server requires communicating $16 * 1024 \approx 16$KB for each position, provided that the continuous parameters are encoded into float16 during communication. As the result, we reduce the communication cost by 1000 times (16 Bits *vs* 16 KB).

Note that previous works (Li and Liang, 2021; Liu et al., 2021) claim that discrete tokens are less expressive than continuous tokens, thus the model capacity may be limited when trained with discrete tokens. However, as described in Section 5.1, datasets of different clients in personalized federated learning may represent different domains/tasks. For such cases, training with continuous prompts via joint training may result in the updated $\boldsymbol{p}_i$ to overfit to the domain/task of client i, causing negative knowledge transfer to other clients when $\boldsymbol{p}_i$ is aggregated with (6). In experiments, we will show that our approach can produce better accuracy compared with joint training with continuous prompt embeddings, while also reducing the communication cost.

4.1.3 Embedding Compression

After the client update, the uploaded $\boldsymbol{p}_i$, for $i = 1 \dots, M$, are aggregated with (6). We can observe that the results $\boldsymbol{p}$ after aggregation can no longer be represented with a single token index, thus cannot be compressed as in Section 4.1.2 when being downloaded to clients. Below we propose to compress $\boldsymbol{p}$ after aggregation with the pretrained token embeddings of the PLM, *i.e.*, estimating $\boldsymbol{p}$ with the matrix of pretrained token embeddings $\boldsymbol{e}(\mathcal{V}) \in \mathbb{R}^{|\mathcal{V}| \times D}$.

This draws from the intuition in previous works on linear word analogies (Ethayarajh et al., 2018; Nissim et al., 2020; Drozd et al., 2016), which show interesting examples with linear operations among the pretrained word/token embeddings, *e.g.*, $\boldsymbol{e}(king) - \boldsymbol{e}(man) + \boldsymbol{e}(woman) \approx \boldsymbol{e}(queen)$ or $\boldsymbol{e}(doctor) - \boldsymbol{e}(man) + \boldsymbol{e}(woman) \approx \boldsymbol{e}(nurse)$. These indicate that a pretrained token embedding can be estimated by a few embeddings of tokens with similar or relevant semantics. As for our $\boldsymbol{p}$, its prompt embeddings is assumed to be within the convex hull of the natural token embeddings. This can be observed from (6), *i.e.*, even $\boldsymbol{p}_i$ that is not updated in client i should also be aggregated from natural token embeddings that appeared as updates in previous rounds. Therefore, it should be viable to estimate $\boldsymbol{p}$ with a few or fixed number of natural token embeddings. For each round of federated learning with aggregated prompt $\boldsymbol{p}$, let $\boldsymbol{p}'$ be the prompts received by the clients from the server after compression in the current round. We denote $\boldsymbol{p}'_{-1}$ as the prompts received by the clients after compression in the previous round. Below, we elaborate on how to compress $\boldsymbol{p}$ into $\boldsymbol{p}'$ for the current update round, given $\boldsymbol{p}'_{-1}$ and $\boldsymbol{e}(\mathcal{V})$.

We should note that different from $\boldsymbol{p}$, the compressed $\boldsymbol{p}'_{-1}$ is accessible by both the server and clients, since it was generated by the server and received by the clients. Thus, instead of directly compressing $\boldsymbol{p}$, we only compress the increment (residual) of $\boldsymbol{p}$ between the previous and current rounds. Specifically, for each position t, we define the residual as $\boldsymbol{R}^t = \boldsymbol{p}^t - \boldsymbol{p}^{t\prime}_{-1}$. For each position t, we want to find a sparse projection from $\boldsymbol{e}(\mathcal{V})$ to $\boldsymbol{R}^t$ so it can be represented/estimated with a limited number of pretrained embeddings. Let $\boldsymbol{I}$ be a sequence of token indices, initialized as $\boldsymbol{I} = [1 \cdots, |\mathcal{V}|]$. We define $\boldsymbol{e}(\mathcal{V})_{\boldsymbol{I}}$ be the rows in $\boldsymbol{e}(\mathcal{V})$ indexed by $\boldsymbol{I}$. Formally, we optimize the following,

$$\boldsymbol{x}^* = \text{argmin}_{\boldsymbol{x}} \, ||\boldsymbol{e}(\mathcal{V})_{\boldsymbol{I}}^T \cdot \boldsymbol{x} - \boldsymbol{R}^t||_2^2 + \alpha||\boldsymbol{x}||_1, \quad (9)$$

$$\boldsymbol{I}_x = \text{argmax}_{|\boldsymbol{I}_x|=L} \sum_{j \in \boldsymbol{I}_x} |\boldsymbol{x}^*[j]|, \; \boldsymbol{I} = \boldsymbol{I}[\boldsymbol{I}_x], \quad (10)$$

where $\boldsymbol{I}[\boldsymbol{I}_x]$ is the value of $\boldsymbol{I}$ indexed by $\boldsymbol{I}_x$. $x \in \mathbb{R}^{|\mathcal{I}| \times 1}$ is the learnt projection, $||\cdot||_1$ and $||\cdot||_2$ are the one and two norms, respectively, and $|\cdot|$ denotes the absolute value. We solve a sparse $\boldsymbol{x}^*$ with LASSO regularization as in (9), with α being the regularization weight. We empirically set $\alpha = 0.2$ for all datasets and clients. $\boldsymbol{x}^*[j]$ is the j^{th} element of $\boldsymbol{x}^*$. Note that (10) takes the top L token indices with the largest absolute projection values in the resulting $\boldsymbol{x}^*$. To minimize the error in estimating $\boldsymbol{R}^t$, the final projection $\boldsymbol{x}_f^* \in \mathbb{R}^{\boldsymbol{I} \times 1}$ is,

$$\boldsymbol{x}_f^* = \mathrm{argmin}_{\boldsymbol{x}_f} \; ||\boldsymbol{e}(\mathcal{V})_{\boldsymbol{I}}^T \cdot \boldsymbol{x}_f - \boldsymbol{R}^t||_2^2. \quad (11)$$

We denote the cardinal of resulting $\boldsymbol{I}$ in (11) as Φ, the number of token embeddings used to approximate $\boldsymbol{R}^t$. Instead of downloading with the aggregated $\boldsymbol{p}$, we download $\{\boldsymbol{I}, \boldsymbol{x}_f^*\}$ to each client. As the result, we only need to download $16 \times 2\Phi$ Bits for each prompt token, consider that both the token index in $\boldsymbol{I}$ and continuous variable in $\boldsymbol{x}_f^*$ are encoded with 16 Bits, as in Section 4.1.2.

The client will reconstruct the residual $\boldsymbol{R}$ via $\hat{\boldsymbol{R}} = \boldsymbol{e}(\mathcal{V})_{\boldsymbol{I}}^T \cdot \boldsymbol{x}_f$ Finally, the compressed prompt received by the clients for the current round is,

$$\boldsymbol{p}^{t'} = \boldsymbol{p}^{t'}_{-1} + \hat{\boldsymbol{R}}^t, \quad (12)$$

$\boldsymbol{p}' = [\boldsymbol{p}^{1'}, \cdots, \boldsymbol{p}^{T'}]$ will be further saved as $\boldsymbol{p}'_{-1}$ for the next round of federated learning. In the experiments, $\boldsymbol{I}$ is selected with two iterations of (9) and (10), as in Algorithm 2.

4.2 Post Tuning

The goal of post tuning is to fine tune the resulting prompt $\boldsymbol{p}$ from the joint training with the local dataset of each client (no communication cost). The resulting $\boldsymbol{p}_i$ should be adapted to the task/domain of client i. Therefore, during post tuning, we adopt the gradient-free method of BBT (Sun et al., 2022b) that allows the prompts being trained in the continuous embedding space. Specifically, for each position t, we follow BBT that reparameterizes $\boldsymbol{p}_i^t$ as,

$$\boldsymbol{p}_i^t = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{p}^t, \quad (13)$$

where $\boldsymbol{z} \in \mathbb{R}^d$, $d << D$, and $\boldsymbol{A} \in \mathcal{R}^{D \times d}$ is a randomly valued fixed matrix that project $\boldsymbol{z}$ into the space of $\boldsymbol{p}^t$. Further, $\boldsymbol{z}$ is the only learnable parameter and is trained with CMA-ES (Hansen and Ostermeier, 2001), a gradient-free method without back-propagation. Please refer to (Sun et al., 2022b) for more details.

5 Experiments

5.1 Experiment Setting

Training: As mentioned above, data from different clients of personalized federated learning may come from different domains/tasks. We experiment with the datasets of FDU-MTL (Liu et al., 2017) and Sentiment140 (Go et al., 2009). FDU-MTL is a domain adaptation dataset for text classification with 16 different domains/clients (each client with a unique domain). We train and evaluate on all the 16 domains. Sentiment140 is a dataset of 1.6 million tweets from 659775 users. We follow (Yan et al., 2020) that treat each user as a client and only keeps clients with more than 40 samples. In experiments, 90% of the clients are sampled as training clients and the rest as testing clients. Please refer to Appendix C for more details.

Evaluation: In addition to the classification accuracy on testing clients, we also evaluate the training efficiency in federated learning. The training efficiency is considered in two perspectives: *1)* Whether the method requires back-propagation, *i.e.*, does the model consumes a large memory footprint for local training? *2)* The communication cost, *i.e.*, the number of communicated Bits between server and clients for each round of federated learning. In calculating the Bits, we assume the token indices are encoded with 16-bit and continuous parameters are converted into float16 during communication, as in Sections 4.1.2 and 4.1.3. Instead of computing the total communicating cost for each round, we calculate the upload and download cost separately, due to the fact that the upload bandwidth is usually smaller than the download bandwidth (Hegedűs et al., 2021), *i.e.*, upload is more expensive than download with the same number of Bits. Another metric for training efficiency for federated learning is the number of floating-point operations, which we discuss in Appendix D.

5.2 Baselines and Our Approaches

All of our baselines are trained with the same model as used in (Sun et al., 2022b). We list the considered baselines are listed as follows: *1) Prompt Tuning* (Li and Liang, 2021), which is to train the separated prompt parameters locally on each testing client with back-propagation. We have learning rate as 1e-2 and batch size 16. *2) Prompt Tuning (Fed)*. The prompts are initially trained with FedAvg (McMahan et al., 2017) on all the clients, then fine tuned on each testing client, as with our

Method	Upload	Download	BP?	Sentiment140	FM(*apparel*)	FM(*mr*)	FM(*baby*)	FM(*books*)	FM(*camera*)	FM(*dvd*)	FM(*electronics*)
Prompt Tuning	0	0	Yes	73.22±14.19	83.42	81.75	79.95	86.38	80.05	86.52	84.18
Prompt Tuning (Fed)	819 KB	819 KB	Yes	74.67±13.28	83.56	81.06	81.05	87.83	81.80	87.96	84.93
Meta Prompt Tuning (Fed)	819 KB	819 KB	Yes	74.89±13.31	82.78	83.35	80.23	88.12	80.34	87.31	84.45
BBT	0	0	No	73.17±14.19	85.93	83.75	81.22	86.10	80.56	85.96	87.76
BBT (Fed)	8 KB	8 KB	No	73.58±13.31	87.44	81.02	82.99	90.19	81.84	87.92	87.74
Ours ($\Phi = 3$)	0.8 KB	4.8 KB	No	74.94±13.46	87.44	80.07	85.53	90.74	82.33	88.48	88.03
Ours ($\Phi = 5$)	0.8 KB	8 KB	No	75.34±12.88	88.54	80.05	86.55	90.21	82.61	88.08	87.78
Ours (FullDownload)	0.8 KB	819 KB	No	76.00±11.98	89.04	81.03	86.78	90.97	83.73	87.18	88.88

Table 1: Results with the Sentiment140 and FDUMLT datasets. For Sentiment140, we report the mean and standard deviation of accuracies on testing clients. For the FDUMLT dataset, we report the accuracies for each of the 16 domains/clients (denoted as FM(*domain name*)) and their average (denoted as FM(*Avg*)). *Upload* and *Download* shows the Bits that is uploaded and downloaded per round of federated learning. *BP?* indicates whether the method requires back-propagation.

Method	FM(*health*)	FM(*imdb*)	FM(*kitchen*)	FM(*magazines*)	FM(*music*)	FM(*software*)	FM(*sports*)	FM(*toys*)	FM(*video*)	FM(*Avg*)
Prompt Tuning	81.98	92.42	82.14	80.68	82.52	83.77	82.41	84.01	82.32	83.41
Prompt Tuning (Fed)	82.74	92.71	83.61	82.97	83.75	84.29	82.89	84.76	82.60	84.28
Meta Prompt Tuning (Fed)	82.34	92.41	84.53	83.25	83.56	83.48	83.58	85.26	82.21	84.20
BBT	84.01	92.13	81.38	81.46	82.28	85.08	82.40	85.53	83.86	84.34
BBT (Fed)	87.06	93.00	85.13	85.90	84.92	84.03	85.46	87.92	85.36	86.12
Ours ($\Phi = 3$)	87.06	92.42	86.73	86.95	85.98	84.55	86.73	87.31	85.91	86.64
Ours ($\Phi = 5$)	87.82	92.71	88.78	87.73	85.19	85.60	86.48	87.31	87.29	87.14
Ours (FullDownload)	89.57	94.27	88.75	87.44	86.34	85.44	87.86	89.31	86.86	87.71

Table 2: Results with the Sentiment140 and FDUMLT datasets (continue).

framework. *3) Meta Prompt Tuning (Fed).* Same as Prompt Tuning (Fed), except that we follow (Fallah et al., 2020) that the prompts are trained using federated meta learning with MAML (Finn et al., 2017). Both *Prompt Tuning (Fed)* and *Meta Prompt Tuning (Fed)* directly communicate all the prompted parameters between server and clients. *4) BBT* (Sun et al., 2022b), train separated prompts locally on each testing client with the gradient-free method of CMA-ES (Hansen and Ostermeier, 2001), as in Section 4.2. This is like the post tuning stage of our approach. *5) BBT (Fed).* Federated training of $\boldsymbol{z}$ in (13) with BBT on training clients and FedAvg on the server. The resulting $\boldsymbol{z}$ is further fine tuned with BBT on the local dataset of each client, *i.e.*, the same as Section 4.2.

In addition, we also implement different variations of our approach: *1) Ours (*Φ*=3 or 5).* We experiment with different values of Φ, controlling the degree of the embedding compression in Section 4.1.3. *2) Ours (FullDownload).* We directly download the aggregated $\boldsymbol{p}$ from (6), without embedding compression. We also discuss the ablation of α in Appendix A.

5.3 Local Search with Different K Values.

As discussed in Section 4.1.2, discrete prompt tokens might be less expressive than continuous prompt embeddings trained with gradients (Li and Liang, 2021; Liu et al., 2021). Thus, one may be concerned about the capability of discrete local search in minimizing the loss functions of different tasks of different clients. From (7), we can observe that such capability is large and determined by the search number K for each step of local search. Ideally, in maximizing the optimization ability of our local search, we can set $K = |\mathcal{V}|$, *i.e.*, and try with the whole vocabulary instead of searching locally. However, such a combinatorial optimization is computationally expensive, thus not compatible with resource constrained clients. There should be a trade-off between the optimization ability and training efficiency for discrete local search.

In this section, we investigate how the optimization ability of our proposed local search is affected by the search number K. In Figure 1, we plot the averaged training loss (4) over all the clients in FDUMTL when training *Ours (*$\Phi = 5$*)* with different K values. We can observe that our local search can effectively minimize the loss function during training. Additionally, we find that the performance gain, *i.e.*, the difference in the optimized loss value, is diminishing when switching from $K = 2$ to $K = 5$ and from $K = 5$ to $K = 8$. However, the introduced computation cost from $K = 2$ to $K = 5$ is the same as that from $K = 5$ to $K = 8$. With such observation, we take $K = 5$ as a trade-off between the computation efficiency and optimization ability, since *1)* local search with $K = 5$ is not very expensive, *e.g.*, comparing the implementation of BBT (Sun et al., 2022b) that requires 20 searches each step. *2)* The performance

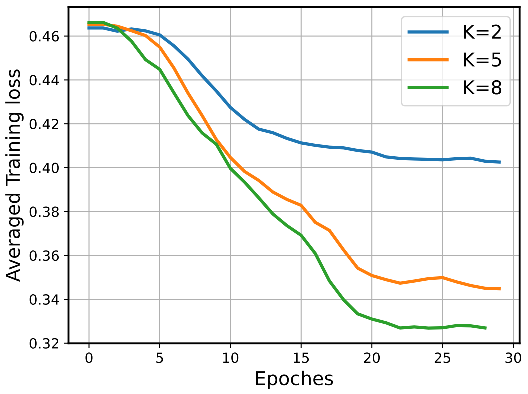


Figure 1: Averaged training loss during joint training of *Ours* ($\Phi = 5$) with different values of K.

gain from $K = 5$ to $K = 8$ is much smaller than that $K = 2$ to $K = 5$, thus increasing the value of K from 5 may not be cost-effective. Therefore, we keep $K = 5$ for all our experiments. Note that such a parameter selection of K only leverages the training data of clients, with no development or testing data involved.

5.4 Results

Table 1 and 2 show the results of personalized federated learning with Sentiment140 and FDU-MTL. Our approaches can achieve highest accuracy, with comparable or much lower communication cost than baselines. This is especially obvious with the upload communication, *i.e.*, the upload cost of our approaches is 10 times smaller than the closest baselines (BBT (Fed)), which thanks to our proposed discrete local search mechanism (Section 4.1.2) that only requires uploading the pretrained token indices to the server. As mentioned in Section 4.2, BBT (Sun et al., 2022b) works by randomly projecting the prompt parameters (with a fixed random matrix $\boldsymbol{A}$) into a small subspace, within which a low-dimensional vector $\boldsymbol{z}$ is trained. However, there is no guarantee that such a random projected subspace can cover directions that captures knowledge that is generalizable across clients. On the contrary, though our local search algorithm is constrained with discrete natural language tokens, such tokens should capture rich semantics of natural language that are expressive enough to describe a pattern that is generalizable across clients. This might explain why our approach of discrete local search with natural language tokens can produce higher accuracy in training with data of different clients.

Additionally, we can observe that compressing using $\Phi = 3$ and $\Phi = 5$ can maintain comparable performance for text classification as with *Ours (FullDownload)*, while substantially decrease the download communication cost. Further, the gradient-based baselines, *i.e.*, those named with prompt tuning, may produce results that is inferior to gradient-free approaches. This may be counter-intuitive since these gradient-based prompt tuning approaches allow training in the whole parameter space of prompt parameters, unlike gradient-free approaches with which the search space for the prompt parameters is usually constrained (Sun et al., 2022b). Thus the learnt continuous prompt embeddings should be more expressive than those from gradient-free approaches, as discussed in Section 4.1.2. However, previous works of gradient-free training with PLMs (Sun et al., 2022b,a) also show results that are better than gradient-free approaches, especially with the scenario of few-shot training. Such a phenomenon may be explained by the over-expressiveness of prompts trained with gradients, *i.e.*, subject to overfitting with limited training data. For the case of federated learning, the prompts trained with gradients may overfit to the task/domain of the clients during local client update, inducing negative knowledge transfer to other clients when being aggregated with 6 in producing model, which is also discussed in Section 4.1.2. Moreover, our implementation of meta prompt learning with MAML (Finn et al., 2017) yields slightly worse results than without meta-learning, *i.e.*, with *Prompt Tuning (Fed)*. We claim that this may not indicate an implementation error, since previous works of federated meta learning (Fallah et al., 2020) also shows that MAML may not always provide improvements compared to meta-learning. For instance, in (Fallah et al., 2020), their meta-learning based method (Per-FedAvg (FO)) can produce inferior results than simple FedAvg (McMahan et al., 2017) in certain scenarios.

6 Conclusion

In this paper, we propose a gradient-free framework that trains with discrete local search on natural language token during personalized federated learning. The discrete local search saves the huge memory consumption caused by back-propagation, while significantly reducing the upload communication cost. We additionally propose a compression mechanism that also reduces the download communication cost of federated learning. Experiments with multiple datasets show that our approach produces superior performance compared with baselines.

References

Atılım Güneş Baydin, Barak A Pearlmutter, Don Syme, Frank Wood, and Philip Torr. 2022. Gradients without backpropagation. *arXiv preprint arXiv:2202.08587*.

Gabriel Belouze. 2022. Optimization without backpropagation. *arXiv preprint arXiv:2209.06302*.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. 2018. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*.

Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. 2015. Binaryconnect: Training deep neural networks with binary weights during propagations. *Advances in neural information processing systems*, 28.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*.

Yunbin Deng. 2019. Deep learning on mobile devices: a review. In *Mobile Multimedia/Image Processing, Security, and Applications 2019*, volume 10993, pages 52–66. SPIE.

Shizhe Diao, Xuechun Li, Yong Lin, Zhichao Huang, and Tong Zhang. 2022. Black-box prompt learning for pre-trained language models. *arXiv preprint arXiv:2201.08531*.

Bo Dong, Yiyi Wang, Hanbo Sun, Yunji Wang, Alireza Hashemi, and Zheng Du. 2022. Cml: A contrastive meta learning method to estimate human label confidence scores and reduce data collection cost. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 35–43.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 3519–3530.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2018. Towards understanding linear word analogies. *arXiv preprint arXiv:1810.04882*.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2020. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Karl Pearson F.R.S. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Yang Gao, Yi-Fan Li, Bo Dong, Yu Lin, and Latifur Khan. 2019. Sim: Open-world multi-task stream classifier with integral similarity metrics. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 751–760. IEEE.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Tao Guo, Song Guo, Junxiao Wang, and Wenchao Xu. 2022. Promptfl: Let federated participants cooperatively learn prompts instead of models–federated learning in age of foundation model. *arXiv preprint arXiv:2208.11625*.

Nikolaus Hansen and Andreas Ostermeier. 2001. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195.

István Hegedűs, Gábor Danner, and Márk Jelasity. 2021. Decentralized learning works: An empirical comparison of gossip learning and federated learning. *Journal of Parallel and Distributed Computing*, 148:109–124.

Bairu Hou, Joe O'Connor, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2022. Promptboosting: Black-box text classification with ten forward passes. *arXiv preprint arXiv:2212.09257*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, pages 4171–4186.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. 2021. A survey on federated learning systems: vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Zhengyang Lit, Shijing Sit, Jianzong Wang, and Jing Xiao. 2022. Federated split bert for heterogeneous text classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Deepak Narayanan, Aaron Harlap, Amar Phanishayee, Vivek Seshadri, Nikhil R Devanur, Gregory R Ganger, Phillip B Gibbons, and Matei Zaharia. 2019. Pipedream: Generalized pipeline parallelism for dnn training. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, pages 1–15.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Yanghua Peng, Yibo Zhu, Yangrui Chen, Yixin Bao, Bairen Yi, Chang Lan, Chuan Wu, and Chuanxiong Guo. 2019. A generic communication scheduler for distributed dnn training acceleration. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, pages 16–29.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2022. Grips: Gradient-free, edit-based instruction search for prompting large language models. *arXiv preprint arXiv:2203.07281*.

Tahseen Rabbani, Brandon Feng, Yifan Yang, Arjun Rajkumar, Amitabh Varshney, and Furong Huang. 2021. Comfetch: Federated learning of large networks on memory-constrained clients via sketching. *arXiv preprint arXiv:2109.08346*.

Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. 2021. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, pages 9489–9502. PMLR.

Tianxiang Sun, Zhengfu He, Hong Qian, Yunhua Zhou, Xuan-Jing Huang, and Xipeng Qiu. 2022a. Bbtv2: Towards a gradient-free future with large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3916–3930.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022b. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, pages 20841–20855. PMLR.

Chaofan Tao, Lu Hou, Wei Zhang, Lifeng Shang, Xin Jiang, Qun Liu, Ping Luo, and Ngai Wong. 2022. Compression of generative pre-trained language models via quantization. *arXiv preprint arXiv:2203.10705*.

Zhuoyi Wang, Yigong Wang, Bo Dong, Sahoo Pracheta, Kevin Hamlen, and Latifur Khan. 2020. Adaptive margin based deep adversarial metric learning. In *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pages 100–108. IEEE.

Yikai Yan, Chaoyue Niu, Yucheng Ding, Zhenzhe Zheng, Fan Wu, Guihai Chen, Shaojie Tang, and Zhihua Wu. 2020. Distributed non-convex optimization with sublinear speedup under intermittent client availability. *arXiv preprint arXiv:2002.07399*.

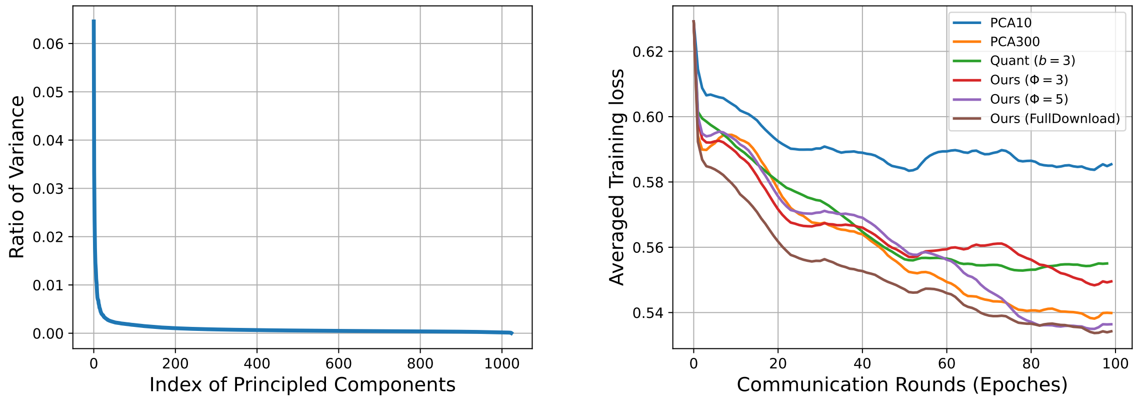
Haodong Zhao, Wei Du, Fangqi Li, Peixuan Li, and Gongshen Liu. 2022. Reduce communication costs and preserve privacy: Prompt tuning method in federated learning. *arXiv preprint arXiv:2208.12268*.

A Ablation study with α

In this section, we conduct an ablation study for the regularization parameter α (default to $\alpha = 0.2$) for the lasso loss in (9). In Table 3, we take Ours ($\Phi = 5$) as an example and report results with $\alpha = 0.2$ (same as in the main paper) and $\alpha = 0$. We can find that the results with $\alpha = 0$ is generally lower than that with $\alpha = 0.2$, indicating the importance of encouraging sparsity with the lassso loss in (9).

B Comparing with PCA compression and quantization

In Section 4.1.3, we present our proposed embedding compression method to reduce the download communication cost. To further validate the effectiveness of the proposed embedding compression,



(a) (b)

Figure 2: (a) The ratio of variance ($\boldsymbol{v}/\sum \boldsymbol{v}_i$) captured by each principled component of the pretrained Roberta-Large Token embeddings. (b) The training loss on Sentiment140 averaged over different clients in each communication round of federated learning for different compression methods. We have the same random seeds and order of training batches for all the methods.

we compare it with the two additionaly baselines: PCA compression and quantization.

PCA Compression: Principled Component Analysis (PCA) (F.R.S., 1901) is a common way of dimensional reduction, *i.e.*, compress the embeddings via representing then with fewer dimensions. Previous works (Cai et al., 2021; Rabbani et al., 2021; Gao et al.) have shown that the learnt token embeddings (contextualized or not) of pre-trained models are distributed in a narrow cone of the embedding space. In another word, the embeddings vectors are generally biased toward the top principled components of learnt embedding matrix. Specially, following the notation of Section 4.1.3, let $\boldsymbol{e}(\mathcal{V}) \in \mathbb{R}^{|\mathcal{V}| \times D}$ be the matrix of pretrained token embeddings. We can compute the principled components of $\boldsymbol{e}(\mathcal{V})$, denoted as,

$$\boldsymbol{E}_c = PCA(\boldsymbol{e}(\mathcal{V})) \tag{14}$$

where each column of $\boldsymbol{E}_c \in \mathbb{R}^{D \times D}$ is a principled component of $\boldsymbol{e}(\mathcal{V})$. We have $\boldsymbol{E}_c^T \cdot \boldsymbol{E}_c = \boldsymbol{I}$, with $\boldsymbol{I} \in \mathbb{R}^{D \times D}$ is the identity matrix. The informativeness of different principled component can be measured by the variance after projecting $\boldsymbol{e}(\mathcal{V})$ onto each of the components,

$$\boldsymbol{v} = \mathrm{Var}(\boldsymbol{e}(\mathcal{V}) \cdot \boldsymbol{E}_c) \tag{15}$$

where Var computes the variance for each row. Assume the index of each component, *i.e.*, the row index of $\boldsymbol{E}_c$, has been ranked by $\boldsymbol{v} = [\boldsymbol{v}_i]_{i=1}^{D}$ (from high to low). We plot the ratio of variance ($\boldsymbol{v}/\sum \boldsymbol{v}_i$) verse the index of each component for Roberta-Large in Figure 2a. We can find that the distribution of $\boldsymbol{e}(\mathcal{V})$ id highly an-isotropic, with much larger variation being captured by the top principled components. Thus, we can represent/compress the aggregated prompt $\boldsymbol{p} \in \mathbb{R}^{T \times D}$ from (6) with the top principled components[2] before downloading it to clients. Specifically, we compress $\boldsymbol{p}$ via,

$$\hat{\boldsymbol{p}} = \boldsymbol{p} \cdot \boldsymbol{E}_c[:n,:]^T \tag{16}$$

where $\hat{\boldsymbol{p}}\mathbb{R}^{T \times n}$ is the compressed prompt and $E_c[:n,:]$ denotes the top-n principled components. After downloading, each client reconstruct $\boldsymbol{p}$ via,

$$\boldsymbol{p} = \hat{\boldsymbol{p}} \cdot \boldsymbol{E}_c[:n,:] \tag{17}$$

In this way, we only need to download n integers (16 bits each) for each prompt token in $\boldsymbol{p}$. The total download bits per communication round is $T \times n \times 16 = 800n$ bits. In comparison with our approach, we experiment with $n = 10$ (denoted as PCA10), so that it has the same download communication cost for each round (8KB) as Ours $\Phi = 5$. We additionally experiment with $n = 300$ (denoted as PCA300), where the prompts are represented by more principled components but also with much larger download communication cost each round (0.24MB).

[2]From Section 4.1.1, each token of p is a convex combination of $\boldsymbol{e}(\mathcal{V})$, thus should also be biased toward (more represented by) the top principled components.

Dataset	Ours ($\Phi = 5, \alpha = 0.2$)	Ours ($\Phi = 5, \alpha = 0$)
FM(*apparel*)	89.04	86.34
FM(*mr*)	81.03	80.32
FM(*baby*)	86.78	84.10
FM(*books*)	90.97	88.25
FM(*camera*)	83.73	81.33
FM(*dvd*)	87.18	87.36
FM(*electronics*)	88.88	87.24
FM(*health*)	89.57	86.8
FM(*imdb*)	94.27	93.51
FM(*kitchen*)	88.75	86.73
FM(*magazines*)	87.44	94.27
FM(*music*)	86.34	85.12
FM(*software*)	85.44	84.31
FM(*sports*)	87.86	84.44
FM(*toys*)	89.31	86.56
FM(*video*)	86.86	86.19
FM(*Avg*)	87.71	85.80
Sentiment140	75.34 ± 12.88	74.35 ± 13.84

Table 3: Ablation study with α.

Method	Upload	Download	BP?	Accuracies
PCA10	0.8KB	8KB	No	73.26 ± 14.77
PCA300	0.8KB	0.24MB	No	75.05 ± 12.69
Quant ($b = 3$)	0.8KB	0.15MB	No	74.44 ± 12.11
Ours ($\Phi = 3$)	0.8KB	4.8KB	No	74.94 ± 13.46
Ours ($\Phi = 5$)	0.8KB	8KB	No	75.34 ± 12.88
Ours (FullDownload)	0.8KB	819KB	No	76.00 ± 11.98

Table 4: Results on Sentiment140 with different compression methods. We report the mean and standard deviation of accuracies on all testing clients.

Quantization: We also compare our approach with quantizing each dimension of $\boldsymbol{p}$ from (6) before downloading. Following previous works (Courbariaux et al., 2015; Tao et al., 2022) of compressing pretrained language models, we quantize each element w of $\boldsymbol{p}$ via,

$$w_q = \beta \cdot Q(clip(w, -\beta, \beta)/\beta) \qquad (18)$$

where Q is a quantization function that maps $clip(w, -\beta, \beta)$ to its closest value in $\{-1, -\frac{k-1}{k}, \cdots, 0, \cdots, \frac{k-1}{k}, 1\}$, $k = 2^{b-1} - 1$. In this way, $Q(clip(w, -\beta, \beta)/\beta)$ can be encoded with b bits. Following (Tao et al., 2022), the scaling factor for each element is shared within the same prompt token embedding. Let $\boldsymbol{p}[i, :]$ be the embedding of the ith prompt token, the scaling factor for each of its element is the maximum absolute value in $\boldsymbol{p}[i, :]$,

$$\beta = max(|\boldsymbol{p}[i, :]|) \qquad (19)$$

Algorithm 2 Compress_Download.

Input: The prompt $\boldsymbol{p}$ without compression, the pretrained embedding matrix $\boldsymbol{e}(\mathcal{V})$.
Output: The reconstructed $\boldsymbol{p}'$.
$\boldsymbol{I} = [1, \cdots, |\mathcal{V}|]$
for $t = 1 \cdots, T$ **do**
  % Embedding compression.
  for $L = [100, 5]$ **do**
    Compute $\boldsymbol{I}$ with (9) and (10).
  end for
  Solve $\boldsymbol{x}_f^*$ with (11).
  % Download.
  Download $\{\boldsymbol{I}, \boldsymbol{x}_f^*\}$ to the clients.
  Compute $\boldsymbol{p}^{t\prime}$ on both server and clients
end for
return $\boldsymbol{p}' = [\boldsymbol{p}^{1\prime}, \cdots, \boldsymbol{p}^{T\prime}]$

For each prompt token with dimension D, we have to download the scaling factor β (16 bits) and b bits for each dimensions, so that the clients can reconstruct w_q. We experiment with $b = 3$, denoted as Quat ($b = 3$). The total download communication cost for each round is $(D \times b + 16) \times T \approx 0.15$MB. Compared with Quat ($b = 3$) that quantizes each dimension of each prompt, our proposed approaches of embedding compression can be regarded as quantizing on the token level, *i.e.*, representing each prompt with pretrained embeddings of discrete tokens.

Results: We report the results with different compress methods in Table 4. We can find that PCA10 has much lower accuracies than Ours ($\Phi = 5$), though sharing the same communication cost. This is because the top 10 principled components cannot capturing enough information about the token embeddings, although the distribution of token embeddings are biased toward the top principled components (Figure 2a). We need to increase the value of n to hundreds in order to get comparable results with our approaches ((*i.e.*, PCA300)), which is at the expense of much higher communication cost. Additionally, we can notice that Quant ($b = 3$) also induces higher download communication cost than our approaches, but yeilding lower accuracies. These results validate the effectiveness of our proposed embedding compression. Additionally, Figure 2b shows the loss values averaged over training clients during federated learning. We can find that our approaches are effective in minimizing the loss function during training (also discussed in

Algorithm 3 Client_Update.

Input: Dataset $\mathcal{D}_i$ for client i, $\boldsymbol{p}'$ from the previous round of communication.
Output: $\boldsymbol{p}_i$ after the client update.
$\boldsymbol{p}_i = \boldsymbol{p}'$
% Training with discrete local search.
for $s = 1 \cdots, S$ **do**
  Randomly sample position t.
  Update $\boldsymbol{p}_i^t$ using (7) and (8) with $\mathcal{D}_i$.
end for
return $\boldsymbol{p}_i$

Section 5.3). We can also find that the final loss values are generally positively correlated with the accuracies in Table 4.

C Additional Explanation

Our model architecture for prompt tuning is the same as in (Sun et al., 2022b). Specifically, the backbone of the PLM is the Roberta-Large model, with $T = 50$ prompt tokens inserted into the input layer. The model is trained with 50 rounds of federated learning for FDUMTL, with each client updated 40 steps for each round. For Sentiment140, we train for 100 rounds and we only sample 50 clients for training during each round (due to the large number of clients in Sentiment140). The implementation of BBT in the both our approaches and the baselines follows (Sun et al., 2022b).

Following previous works of gradient-free learning (Sun et al., 2022b; Hou et al., 2022), we consider the few-shot scenario for each testing client. Specifically, we assume there are 16 samples for each class in each testing client during post-tuning. For FDUMTL, these datasets are sampled from the development split in each domain. For sentiment140, these are sampled from the datasets of each testing client, with the rest data of each client used for testing after post tuning. We additionally sample a development dataset (not overlapped with data for training) from the development split for each client for FDUMLT with the same size as the training set, since development datasets are also used in previous works of gradient-free training (Sun et al., 2022b; Hou et al., 2022). We evaluate the classification accuracy of the resulting models on the test set of each client, averaged over four random seeds. We do not sample development datasets for Sentiment140 since no development datasets are provided. Note that our experiments are based only on English datasets and it would also be interesting for future works studying multilingual federated learning. We provide the algorithm for Client_Update and Compress_ownload in Algorithm 3 and 2, respectively.

D The number of floating-point operations during federated learning

From the previous work (Sun et al., 2022b) of gradient-free training for PLMs, the number of floating-point operations with gradient-free training can be evaluated via the number of model queries (i.e., how many times a model is forwarded). For all the methods in the paper, we have the same number of communication rounds and same number of update steps for each client per round. Thus, the number of floating-point operations is proportional to the number of model queries per step when training on each client. We keep all the discussed approaches with the proposed discrete local search method having 5 model queries per step (i.e., $K = 5$ as in Section 5.3), including the approaches denotes with "Ours" and those in Appendix B. Thus, all these approaches have the same number of model queries during federated learning. Comparably, our gradient-free federated learning baseline (i.e., BBT(Fed), there was no previous works on gradient-free federated learning with pretrained models) have 20 model queries per step, following the original implementation of (Sun et al., 2022b). This implies that our methods (5 queries per step) only use $1/4$ $(5/20)$ times of floating-point operations during federated learning, while having better performance than BBT(Fed). Since we target the scenario that clients has limited memory access, where back-propagation might not be viable (Section 1), we mostly compare the number of floating-point operations of our methods with gradient-free federated learning baselines. Provided the number of floating-point operations during federated learning, the training efficiency can be further enhenced by system designs, *e.g.*, the parallelism strategy (Narayanan et al., 2019) or communication scheduler (Peng et al., 2019), which are out of the scope of this paper.

E Overhead

Our way of converting the prompt token index of each position to 16 bits (Section 5.1) induces no computational overhead, if we save the 16 bits index for each position during training (50 prompt positions in total, *i.e.*, $T = 50$). The uploading of such bits is the same as uploading any model parameters in federated learning. There is not need of additionaly designed software implementation. Actually, by only uploading 16 bits for each position, we save the upload time compared with uploading the prompy embedding (the gradient-based methods in Table 1 and 2).