FedKD: Communication Efficient Federated Learning via Knowledge Distillation

Chuhan Wu¹ Fangzhao Wu² Lingjuan Lyu³ Yongfeng Huang¹ Xing Xie²

¹Department of Electronic Engineering & BNRist, Tsinghua University

²Microsoft Research Asia, ³Sony AI

{wuchuhan15, wufangzhao, lingjuanlvsmile}@gmail.com yfhuang@tsinghua.edu.cn, xingx@microsoft.com

Abstract

Federated learning is widely used to learn intelligent models from decentralized data. In federated learning, clients need to communicate their local model updates in each iteration of model learning. However, model updates are large in size if the model contains numerous parameters, and there usually needs many rounds of communication until model converges. Thus, the communication cost in federated learning can be quite heavy. In this paper, we propose a communication efficient federated learning method based on knowledge distillation. Instead of directly communicating the large models between clients and server, we propose an adaptive mutual distillation framework to reciprocally learn a student and a teacher model on each client, where only the student model is shared by different clients and updated collaboratively to reduce the communication cost. Both the teacher and student on each client are learned on its local data and the knowledge distilled from each other, where their distillation intensities are controlled by their prediction quality. To further reduce the communication cost, we propose a dynamic gradient approximation method based on singular value decomposition to approximate the exchanged gradients with dynamic precision. Extensive experiments on benchmark datasets in different tasks show that our approach can effectively reduce the communication cost and achieve competitive results.

1 Introduction

Privacy protection of user data is a very important issue (Shokri and Shmatikov, 2015). Federated learning is a well-known technique to learn intelligent models from decentralized user data (McMahan et al., 2017). It has been widely used in various applications such as intelligent keyboard (Hard et al., 2018), personalized recommendation (Qi et al., 2020) and topic modeling (Jiang et al., 2019).

In federated learning, the private data is locally stored on different clients (Yang et al., 2019). Each

client keeps a local model and computes the model updates from its local data. In each iteration, the model updates from a number of clients are uploaded to a server, which aggregates the local model updates into a global one to update its maintained global model. Then, the server distributes the global update to each client to conduct a local model update. This process is iteratively executed for many rounds until the model converges. In this framework, the server and clients need to intensively communicate the model updates. However, the communication cost is enormous if the model is in large size, which hinders the applications of many powerful but large-scale models like BERT (Devlin et al., 2019) to federated learning.

In this paper, we propose a communication efficient federated learning method based on knowledge distillation (FedKD). Instead of directly communicating the large models between the clients and server, in FedKD a small student model and a large teacher model are distilled from each other, where only the student model is shared by different clients and learned collaboratively, which can effectively reduce the communication cost. More specifically, each client maintains a large local teacher model and a local copy of a small student model that is shared among different clients. We propose an adaptive knowledge distillation method to enable the local teacher and student to learn from both the local data on its client and the knowledge distilled from each other, where their distillation intensities are controlled by the correctness of their predictions. The local teacher model on each client is locally updated, while the local updates of the student models from different clients are uploaded to a central server, which aggregates these local updates into a global one. The server further distributes the global update to different clients to update their local student models. This process is iteratively executed until the student model converges. In addition, to further reduce the communication cost when exchanging the student model updates, we propose a dynamic gradient approximation method based on singular value decomposition (SVD) to compress the communicated gradients with dynamic precision. Extensive experiments on benchmark datasets for different tasks validate that our approach can effectively reduce communication costs in federated learning and meanwhile achieve competitive performance.

The contributions of this paper are as follows:

- We propose a communication efficient federated learning approach based on knowledge distillation, which can achieve competitive results with much less communication cost.
- We propose an adaptive mutual knowledge distillation method to encourage teacher and student to learn from each other and be aware of their prediction correctness.
- We propose a dynamic gradient approximation method based on SVD for gradient compression with dynamic precision to further reduce communication cost.
- We conduct extensive experiments on benchmark datasets for different tasks to verify the effectiveness and efficiency of our approach.

2 Related Work

2.1 Federated Learning

Federated learning (FL) (McMahan et al., 2017) is a privacy-aware technique to learn intelligent models from decentralized data storage, where the raw user data never leaves where it is stored. It has been widely used in many applications like intelligent keyboard (Hard et al., 2018), personalized recommendation (Lin et al., 2020a; Qi et al., 2020), topic modeling (Jiang et al., 2019) and medical natural language processing (Ge et al., 2020). In federated learning, there are usually a number of user devices that locally keep the privacy-sensitive user data, and a server that coordinates these user devices for collaborative model learning. Each user device contains a local model copy and computes the model update based on the local data. The model updates from a certain number of user devices are uploaded to the server, which aggregates the local updates into a global one for updating its maintained global model. The updated global model is further distributed to user devices to update their local model copies. This process will

be repeated until the model is fully trained. Since the model updates usually contain much less private information (McMahan et al., 2017), federated learning can exploit decentralized data for model learning and significantly reduce privacy and security risks. However, since model updates are communicated between the server and user clients for many rounds, the communication cost would be huge if the model is large. To remedy this issue, we propose a communication efficient federated learning method with knowledge distillation, which can reduce the parameters to be communicated and meanwhile keep competitive model performance.

2.2 Knowledge Distillation

Knowledge distillation is a technique to transfer knowledge from a large teacher model (e.g., BERT) to a small student model (Hinton et al., 2015), which is widely used for model compression (Sanh et al., 2019; Sun et al., 2019; Jiao et al., 2020; Wang et al., 2020b). For example, Sanh et al. (2019) proposed a DistilBERT approach that distills useful knowledge from the output using the distillation loss and the hidden states of the teacher model via a cosine loss. Sun et al. (2019) proposed a BERT-PKD approach that aligns the hidden states of the student model with the teacher using a mean squared error loss. Jiao et al. (2020) proposed a TinyBERT approach that can additionally transfer useful knowledge from the attention matrix of the teacher model. However, these methods usually require centralized data storage, which may pose privacy issues during data collection.

2.3 Communication Efficient FL

Generally, the communication efficiency of federated learning can be improved by gradient compression (Konečný et al., 2016; Caldas et al., 2018; Rothchild et al., 2020) and knowledge distillation (Sui et al., 2020). Both genres of methods are orthogonal in reducing the communication cost of federated learning and are usually compatible with each other. A core technique used by existing knowledge distillation-based federated learning methods is codistillation (Anil et al., 2018). In this method, the models on different clients are learned on the same dataset. The output of each model is regularized to be similar to the ensemble of predictions from all models via a distillation loss. The idea of codistillation is used by several methods to reduce communication cost of federated learning (Sui et al., 2020; Li and Wang, 2019; Seo et al.,

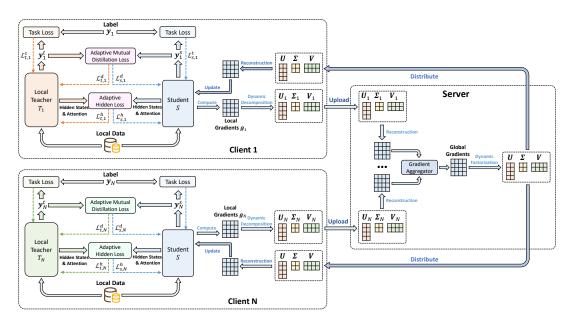


Figure 1: The framework of our *FedKD* approach.

2020; Lin et al., 2020b; Sun and Lyu, 2020). For example, Sui et al. (2020) proposed a federated ensemble distillation approach for medical relation extraction. It first learns student models locally on each client and then uses the student models to generate predictions on a shared dataset and upload them to a server. The server ensembles the predictions from different clients as a virtual teacher and computes the distillation loss between the teacher and students. In this way, the model parameters do not need to be uploaded, and only the predictions on the shared dataset are communicated, which can reduce the communication cost. However, these methods require a dataset that is shared among different clients to conduct ensemble distillation. Unfortunately, in many scenarios such as personalized recommendation, the data (e.g., user behavior logs) is highly privacy-sensitive and cannot be shared or exchange among different clients. Thus, these methods cannot be applied to these scenarios. By contrast, our approach circumvents the need of a shared dataset because the teacher models in our approach are locally stored on different clients. Our approach can also effectively reduce the communication cost by communicating a distilled tiny student model instead of the original large model and using SVD to reduce gradient size.

3 FedKD

In this section, we introduce our communication efficient federated learning approach based on knowledge distillation (FedKD). We first present a def-

inition of the problem studied in this paper, then introduce the details of our approach, and finally present some discussions on the computation and communication complexity of our approach.

3.1 Problem Definition

In our approach, we assume that there are N clients that locally store their private data, where the raw data never leaves the client where it is stored. We denote the dataset on the i-th client as D_i . In our approach, each client keeps a large local teacher model T_i with a parameter set Θ_i^t and a local copy of a smaller shared student model S with a parameter set S. In addition, a central server coordinates these clients for collaborative model learning. The goal is to learn a strong model in a privacy-preserving way with less communication cost.

3.2 Federated Knowledge Distillation

Next, we introduce the details of our federated knowledge distillation framework, as shown in Figure 1. In each iteration, each client simultaneously computes the update of the local teacher model and the student model based on the supervision of the labeled local data and the knowledge distilled from each other. The teacher models are locally updated, while the student model is shared among different clients and are learned collaboratively. Since the local teacher models have more sophisticated architectures than the student model, the useful knowledge encoded by the teacher model can help teach the student model. In addition, since the

teacher model can only learn from local data while the student model can see the data on all clients, the teacher can also benefit from the knowledge distilled from the student model.

In our approach, we use three loss functions to learn student and teacher models locally, including an adaptive mutual distillation loss to transfer knowledge from output soft labels, an adaptive hidden loss to distill knowledge from the hidden states and self-attention heatmaps, and a task loss to directly provide task-specific supervision for learning the teacher and student models. We denote the soft probabilities of a sample x_i predicted by the local teacher and student on the *i*-th client as \mathbf{y}_{i}^{t} and \mathbf{y}_{i}^{s} , respectively. Since incorrect predictions from the teacher/student model may mislead the other one in the knowledge transfer, we propose an adaptive method to weight the distillation loss according to the quality of predicted soft labels. We first use the task labels to compute the task losses for the teacher and student models (denoted as $\mathcal{L}_{t,i}^t$ and $\mathcal{L}_{s,i}^{s}$). We denote the gold label of x_i as \mathbf{y}_i , and the task losses are formulated as follows:

$$\mathcal{L}_{t,i}^{t} = CE(\mathbf{y}_{i}, \mathbf{y}_{i}^{t}), \tag{1}$$

$$\mathcal{L}_{s,i}^{t} = CE(\mathbf{y}_{i}, \mathbf{y}_{i}^{s}), \tag{2}$$

where CE stands for cross-entropy. The adaptive distillation losses for both teacher and student models (denoted as $\mathcal{L}^d_{t,i}$ and $\mathcal{L}^d_{s,i}$)are formulated as follows:

$$\mathcal{L}_{t,i}^{d} = \frac{\mathrm{KL}(\mathbf{y}_{i}^{\mathrm{s}}, \mathbf{y}_{i}^{\mathrm{t}})}{\mathcal{L}_{t,i}^{t} + \mathcal{L}_{s,i}^{t}},\tag{3}$$

$$\mathcal{L}_{s,i}^{d} = \frac{\mathrm{KL}(\mathbf{y}_{i}^{t}, \mathbf{y}_{i}^{s})}{\mathcal{L}_{t,i}^{t} + \mathcal{L}_{s,i}^{t}},$$
(4)

where KL means the Kullback–Leibler divergence. In this way, the distillation intensity is weak if the predictions of teacher and student are not reliable. The distillation loss becomes dominant if the student and teacher are well tuned, which has the potential to mitigate the risk of overfitting. In addition, previous works have validated that transferring knowledge between the hidden states (Sun et al., 2019) and hidden attention matrices (Jiao et al., 2020) (if available) is beneficial for student teaching. Thus, taking language model distillation as an example, we also introduce additional adaptive hidden losses to align the hidden states and attention heatmaps of the student and the local teachers. The losses for the teacher and student

models (denoted as $\mathcal{L}_{t,i}^h$ and $\mathcal{L}_{s,i}^h$) are formulated as follows:

$$\mathcal{L}_{t,i}^{h} = \mathcal{L}_{s,i}^{h} = \frac{\text{MSE}(\mathbf{H}_{i}^{t}, \mathbf{W}_{i}^{h} \mathbf{H}^{s}) + \text{MSE}(\mathbf{A}_{i}^{t}, \mathbf{A}^{s})}{\mathcal{L}_{t,i}^{t} + \mathcal{L}_{s,i}^{t}},$$
(5)

where MSE stands for the mean squared error, \mathbf{H}_{i}^{t} , \mathbf{A}_{i}^{t} , \mathbf{H}^{s} , and \mathbf{A}^{s} respectively denote the hidden states and attention heatmaps in the i-th local teacher and the student, and \mathbf{W}_{i}^{h} is a learnable linear transformation matrix. Here we propose to control the intensity of the adaptive hidden loss based on the prediction correctness of the student and teacher. Besides, motivated by the task-specific distillation framework in (Tang et al., 2019), we also learn the student model based on the task-specific labels on each client. Thus, on each client the unified loss functions for computing the local update of teacher and student models (denoted as $\mathcal{L}_{t,i}$ and $\mathcal{L}_{s,i}$) are formulated as follows:

$$\mathcal{L}_{t,i} = \mathcal{L}_{t,i}^d + \mathcal{L}_{t,i}^h + \mathcal{L}_{t,i}^t, \tag{6}$$

$$\mathcal{L}_{s,i} = \mathcal{L}_{s,i}^d + \mathcal{L}_{s,i}^h + \mathcal{L}_{s,i}^t, \tag{7}$$

The student model gradients \mathbf{g}_i on the *i*-th client can be derived from $\mathcal{L}_{s,i}$ via $\mathbf{g}_i = \frac{\partial \mathcal{L}_{s,i}}{\partial \Theta^s}$, where Θ^s is the parameter set of student model. The local teacher model on each client is immediately updated by their local gradients derived from the loss function $\mathcal{L}_{t,i}$.

Afterwards, the local gradients g_i on each client will be uploaded to the central server for global aggregation. Since the raw model gradients may still contain some private information (Zhu and Han, 2020), we encrypt the local gradients before uploading. The server receives the local student model gradients from different clients and uses a gradient aggregator¹ to synthesize the local gradients into a global one (denoted as g). The server further delivers the aggregated global gradients to each client for a local update. The client decrypts the global gradients to update its local copy of the student model. This process will be repeated until both student and teacher models converge. Note that in the test phase, the teacher model is used for label inference.

3.3 Dynamic Gradients Approximation

In our *FedKD* framework, although the size of student model updates is smaller than the teacher models, the communication cost can still be relatively

¹We use the FedAvg method for simplicity.

high when the student model is not tiny. Thus, we propose to a dynamic gradients approximation method to compress the gradients exchanged among the server and clients to further reduce computational cost. As shown in Fig. 1, we first factorize the local gradients into smaller matrices before uploading them. The server reconstructs the local gradients by multiplying the factorized matrices before aggregation. The aggregated global gradients are further factorized, which are distributed to the clients for reconstruction and model update. More specifically, we denote the gradient $\mathbf{g}_i \in \mathbb{R}^{P \times Q}$ as a matrix with P rows and Q columns (we assume $P \geq Q$).² It is approximately factorized into the multiplication of three matrix, i.e., $\mathbf{g}_i \approx \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i$, where $\mathbf{U}_i \in \mathbb{R}^{P \times K}$, $\mathbf{\Sigma}_i \in \mathbb{R}^{K \times K}$, $\mathbf{V}_i \in \mathbb{R}^{K \times Q}$ are factorized matrices and K is the number of retained singular values. If the value of K satisfies $PK + K^2 + KQ < PQ$, the size of uploaded and downloaded gradients can be reduced. We denote the singular values of \mathbf{g}_i as $[\sigma_1, \sigma_2, ..., \sigma_O]$ (ordered by their absolute values). To control the approximation error, we use an energy threshold Tto decide how many singular values are kept, which is formulated as follows:

$$\min_{K} \frac{\sum_{i=1}^{K} \sigma_{i}^{2}}{\sum_{i=1}^{Q} \sigma_{i}^{2}} > T.$$
 (8)

To better help the model converge, we propose to use a dynamic value of T. The function between the threshold T and the percentage of training steps t is formulated as follows:

$$T(t) = T_{start} + (T_{end} - T_{start})t, t \in [0, 1], (9)$$

where T_{start} and T_{end} are two hyperparameters that control the start and end values of T. In this way, the student model is learned on roughly approximated gradients at the beginning, while learned on more accurately approximated gradients when the model gets to convergence, which can help learn a more accurate student model.

To help readers better understand how *FedKD* works, we summarize the entire workflow of *FedKD* in the Algorithm 1 in Appendix.

3.4 Complexity Analysis

In this section, we will present some analysis on the complexity of our *FedKD* approach in terms of computation and communication cost. We denote the number of communication rounds as Rand the average size of dataset on each client as D. Thus, the computational cost of directly learning a large model (the parameter set is denoted as Θ^t) in a federated way is $O(RD|\Theta^t|)$, and the communication cost is $O(R|\Theta^t|)$.³ In *FedKD*, the communication cost is $O(R|\Theta^s|/\rho)$ (ρ is the gradient compression ratio), which is much smaller because $|\Theta^s| \ll |\Theta^t|$ and $\rho > 1$. The computational cost contains three parts, i.e., local teacher model learning, student model learning and gradient compression/reconstruction, which are $O(RD|\Theta^t|)$, $O(RD|\Theta^s|)$ and $O(RPQ^2)$, respectively. The total computational cost of FedKD is $O(RD|\Theta^t| + RD|\Theta^s| + RPQ^2)$. In practice, compared with the standard FedAvg (McMahan et al., 2017) method, the extra computational cost of learning the student model in FedKD is much smaller than learning the large teacher model, and SVD can also be very efficiently computed in parallel. Thus, FedKD is efficient in terms of both communication and computation.

4 Experiments

4.1 Datasets and Experimental Settings

Our experiments are conducted in two tasks that involve user data. The first one is personalized news recommendation, which needs to predict whether a user will click a candidate news based on the user interest inferred from historical news click behaviors. In this task we use the MIND (Wu et al., 2020) dataset.⁴ It contains the news impression logs of 1 million users on the Microsoft News platform during 6 weeks. The logs in the last week are used for test, and the rest are for training and validation. The second one is adverse drug reaction (ADR) mentioning tweet detection, which is a binary classification task. We use the dataset released by the 3rd shared task of the SMM4H 2018 workshop (Weissenbacher et al., 2018).⁵ We denote this dataset as SMM4H. The original SMM4H dataset contains 25,678 tweet IDs. However, since many tweet texts in this dataset are no longer available, we only crawled 16,694 tweets for experiments. Following (Wu et al., 2019d), we use 80% of the dataset for training, 10% for validation and 10% for test. The detailed statistics of these two datasets

²We formulate \mathbf{g}_i as a single matrix for simplicity. In practice, different parameter matrices in the model are factorized independently. The global gradients on the server are factorized in the same way.

³We assume the cost is linearly proportional to model sizes.

⁴https://msnews.github.io/

⁵https://healthlanguageprocessing.org/smm4h18

are summarized in Table 1. To simulate the scenario where private data is decentralized on different clients, we randomly divide the training data into 4 folds and assume that each fold is locally stored on different clients.

MIND					
# users	1,000,000	# impressions 15,777,			
# news	161,013	# clicks	24,155,470		
avg. title len.	11.52	# training samples	2,186,683		
# validation samples	365,200	# test samples	2,341,619		
SMM4H					
# tweets	16,694	4 # positives			
avg. tweet len.	16.48	# negatives 15,3			

Table 1: Statistics of the datasets.

In our experiments, on each client we use the UniLM-Base (Bao et al., 2020) model as the local teacher.⁶ We use its submodels with the first 4 or 2 Transformer layers as the student models. On the MIND dataset, we incorporate the language model as the news encoder of NAML. On the SMM4H dataset we apply an attentive pooling and a dense layer after the language model for text classification. The energy thresholds T_{start} and T_{end} are 0.95 and 0.98, respectively. The optimizer we use is Adam (Bengio and LeCun, 2015). 7 Following (Wu et al., 2020), on the MIND dataset, we use AUC, MRR, nDCG@5 and nDCG@10 as the metrics. On the SMM4H dataset, we use precision, recall and Fscore of the positive class as the metrics (Wu et al., 2019d). We repeat each experiment repeat 5 times to mitigate occasionality.

4.2 Performance Evaluation

First, we compare the performance and communication \cos^8 of FedKD with several additional baselines, including: (1) UniLM (Local), learning the full UniLM model with the local data on a client; (2) UniLM (Cen), learning the full UniLM model on centralized data; (3) UniLM (Fed), learning the full UniLM model in the standard federated framework; (4) DistilBERT (Sanh et al., 2019), finetuning the DistilBERT model in federated learning; (5) BERT-PKD (Sun et al., 2019), finetuning BERT-PKD in a federated manner; (6) Tiny-BERT (Jiao et al., 2020), finetuning TinyBERT in a federated way; (7) $UniLM_{4/2}$, using the first 4 or 2 layers of UniLM in federated learning. (8)

Methods	AUC	MRR	nDCG@5	nDCG@10	Comm. Cost per Client
UniLM (Local)	68.8±0.5	33.5±0.4	36.6±0.5	42.4±0.6	-
UniLM (Cen)	71.0 ± 0.1	35.8 ± 0.1	39.0 ± 0.1	44.8 ± 0.1	-
UniLM (Fed)	70.9 ± 0.3	35.7 ± 0.2	38.9 ± 0.3	44.7 ± 0.4	2.05GB
DistilBERT ₆	69.3±0.2	34.0±0.2	37.5±0.2	43.0±0.1	1.03GB
DistilBERT ₄	69.0 ± 0.2	33.7 ± 0.1	37.0 ± 0.1	42.6 ± 0.2	0.69GB
BERT-PKD $_6$	69.6 ± 0.2	34.4 ± 0.3	37.7 ± 0.3	43.4 ± 0.2	1.03GB
BERT-PKD $_4$	69.2 ± 0.2	33.8 ± 0.2	37.1 ± 0.3	42.9 ± 0.3	0.69GB
TinyBERT ₆	69.7 ± 0.2	34.5 ± 0.2	37.9 ± 0.1	43.5 ± 0.2	1.03GB
$TinyBERT_4$	69.4 ± 0.3	33.9 ± 0.3	37.5 ± 0.2	43.1 ± 0.2	0.17GB
UniLM ₄	69.6±0.1	34.4±0.2	37.7±0.1	43.4±0.2	0.69GB
UniLM $_2$	68.9 ± 0.2	33.6 ± 0.2	36.8 ± 0.2	42.5 ± 0.1	0.35GB
FetchSGD	70.5±0.4	35.2±0.3	38.2±0.3	44.0±0.4	0.51GB
FedDropout	70.5 ± 0.2	35.1 ± 0.2	38.3 ± 0.3	44.2 ± 0.3	1.23GB
FedKD ₄	71.0 ±0.1	35.6±0.1	38.9±0.1	44.8 ±0.1	0.19GB
$FedKD_2$	70.5 ± 0.1	35.3 ± 0.2	38.6 ± 0.1	44.3 ± 0.2	0.11GB

Table 2: Performance of different methods on MIND.

Methods	Precision	Recall Fscore	Comm. Cost	
Methous	Frecision	Kecan	rscore	per Client
UniLM (Local)	53.2±1.3	54.6±1.4	53.9±1.1	-
UniLM (Cen)	60.3 \pm 0.7	61.6 ± 0.8	60.8 \pm 0.4	-
UniLM (Fed)	59.1 ± 0.6	62.3 ± 0.6	$60.6 {\pm} 0.4$	1.37GB
DistilBERT ₆	56.8 ± 0.8	59.2±0.8	57.9±0.5	0.69GB
DistilBERT ₄	56.5 ± 0.9	58.4 ± 1.1	57.1 ± 0.7	0.46GB
BERT-PKD $_6$	56.9 ± 0.9	$60.4 {\pm} 0.8$	58.4 ± 0.6	0.69GB
BERT-PKD $_4$	56.3 ± 1.1	59.9 ± 0.7	58.0 ± 0.6	0.46GB
TinyBERT ₆	57.4 ± 0.8	60.5 ± 0.6	58.6 ± 0.5	0.69GB
$TinyBERT_4$	57.0 ± 0.7	59.9 ± 1.2	58.3 ± 0.7	0.12GB
UniLM ₄	56.1 ± 0.9	60.6±0.9	58.2±0.5	0.46GB
$UniLM_2$	53.8 ± 0.8	59.1 ± 1.0	56.3 ± 0.6	0.24GB
FetchSGD	57.5±0.9	60.4±1.1	59.0±0.8	0.34GB
FedDropout	57.8 ± 1.0	61.0 ± 0.8	59.4 ± 0.6	0.82GB
FedKD ₄	59.4±0.6	62.8 ±0.9	60.7±0.5	0.12GB
$FedKD_2$	58.2 ± 0.7	62.4 ± 0.9	59.8 ± 0.6	0.07GB

Table 3: Performance of different methods on SMM4H.

FetchSGD (Rothchild et al., 2020), a count sketch based communication efficient federated learning method. (9) FedDropout (Caldas et al., 2018), a federated dropout method to reduce the number of exchanged parameters. In the methods (4)-(6), we compare the performance of their officially released 6-layer and 4-layer models. In methods (8) and (9), we use the full UniLM model. The results on the MIND and SMM4H datasets are respectively shown in Tables 2 and 3. From the results, we have the following findings. First, compared with UniLM (local), other methods achieve better performance. This is because the local data on a single client may not be sufficient to learn a strong model, while federated learning can exploit data decentralized on multiple clients to facilitate model training. Second, although UniLM achieves the best performance, the communication cost for model learning is huge (e.g., over 2GB for each client on the MIND dataset). Thus, it may be difficult to incorporate it in real-world applications. Third, compared with the off-the-shelf distilled models like *DistilBERT*, BERT-PKD and TinyBERT, our FedKD approach performs better. This is because the former meth-

⁶We take pre-trained language model distillation as a representative example in our experiments.

⁷The detailed hyperparameter settings of our approach and baselines are in the Appendix.

⁸The communication cost on the two datasets are slightly different due to the number of updated token embeddings.

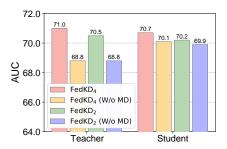


Figure 2: Influence of mutual distillation on the student and teacher models.

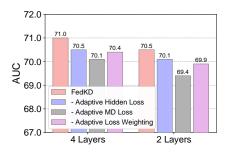


Figure 3: Effect of adaptive mutual distillation.

ods are distilled in a task-agnostic manner, which may be suboptimal in downstream tasks without further task-specific distillation. Fourth, FedKD also outperforms $UniLM_4$ and $UniLM_2$. This is because FedKD can learn useful knowledge from the output and intermediate results of the sophisticated local teacher models while *UniLM*₄ and *UniLM*₂ cannot. Fifth, FedKD can achieve better performance and lower communication cost than other communication efficient methods like FedtchSGD and FedDropout. It is because FedKD can transfer rich knowledge between the teacher and student models to improve the model performance, and can reduce the communication cost by exchanging the updates of a small student model and meanwhile compress the gradients with SVD. Sixth, the communication cost of FedKD is much less than the original *UniLM* model, and the performance of FedKD is comparable with UniLM (Fed) and *UniLM (Cen)*. These results show that *FedKD* can effectively reduce the communication cost of federated learning while keeping good performance.

4.3 Effectiveness of Adaptive Mutual Distillation

We also verify the effectiveness of our proposed adaptive mutual distillation method. We first compare the performance of *FedKD* models trained with or without mutual distillation (the teacher

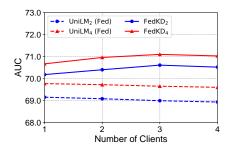


Figure 4: Influence of client number.

model is only learned on local data), as shown in Fig. 2.9 We observe that mutual distillation can effectively improve the performance of both teacher and student models with different sizes, especially the teacher model. This is because useful knowledge transferred between the teachers and student can help student better imitate the complicated teacher models, and can help teachers break the limitation of the amount of local labeled data. In addition, we observe that local teachers slightly outperform the student. Thus, we choose to use the teacher models for inference in the test stages.

We further compare *FedKD* and its variants with the adaptive mutual distillation loss, the adaptive hidden loss or the adaptive loss weighting method removed, as shown in Fig. 3 (we report the performance of teacher models). We can see that both adaptive mutual distillation and adaptive hidden losses are useful for improving the model performance. In addition, the performance is suboptimal when the adaptive loss weighting method is removed (this variant is similar to the standard mutual distillation (Zhang et al., 2018)). This is because weighting the distillation and hidden losses can be aware of the correctness of model predictions, which may help distill higher-quality knowledge and meanwhile mitigate the risk of overfitting.

4.4 Influence of Client Number

We study the influence of client number on the model performance in this section. We divide the full training data into different numbers of folds to simulate the scenarios with different amounts of labeled data on each client. Figure 4 shows the performance of FedKD and $UniLM_{4/2}$ under different numbers of clients. We find the performance of FedKD is similar and can even be slightly improved when more clients are involved. This is because FedKD can learn from multiple teacher models on

⁹We only include results on *MIND* due to space limit. The results on *SMM4H* are in Appendix.

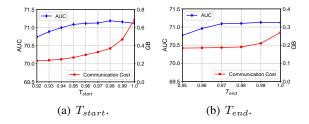


Figure 5: Influence of T_{start} and T_{end} on model performance and communication cost.

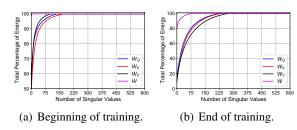


Figure 6: Cumulative energy distributions of singular values of different parameter gradient matrices. W_Q : query parameters, W_K : key parameters, W_V : value parameters, W: feed-forward network parameters.

different clients, which can encode richer knowledge when more teacher models participate. On the contrary, the performance of $UniLM_{4/2}$ (Fed) slightly declines with the increase of client number. This may be because the vanilla FedAvg method has some performance sacrifice by learning models for multiple epochs on limited local data.

4.5 Impact of Energy Threshold

We then study the influence of the energy threshold T_{start} and T_{end} on the performance and communication cost of our approach. We first vary T_{start} under $T_{end}=1$, and the results are shown in Fig. 6(a). We find the communication cost is smaller when T_{start} is smaller, while we observe that the performance starts to drop quickly when $T_{start}<0.95$. Thus, we chose $T_{start}=0.95$ to balance communication cost and model performance. Under $T_{start}=0.95$, we then vary T_{end} to compare the performance and communication cost, as shown in Fig. 6(b). In a similar way, we choose $T_{end}=0.98$ to achieve a good tradeoff between model accuracy and communication cost.

4.6 Analysis of Dynamic Gradient Approximation

Finally, we present some analysis of our proposed SVD-based gradient compression method. We show the cumulative energy distributions of singu-

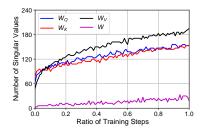


Figure 7: Evolution of the number of required singular values under T=0.95.

lar values of different parameter gradient matrices in the UniLM model in Fig. 6, which reveals several interesting findings. First, all kinds of parameter matrices in UniLM are low-rank, especially the parameters in the feed-forward network. Thus, the communication cost can be greatly reduced by compressing the low-rank gradient matrices. In addition, we find the singular value energy is more concentrated at the beginning than the end of training. This may be because when the model is not well-tuned, the gradients may have more low frequency components that aim to push the model to converge more quickly. However, when the model gets to converge, the updates of model parameters are usually subtle, which yields more high frequency components. The evolution of required singular values under T = 0.95 is shown in Fig. 7. We can see that more singular values need to be retained to achieve the same energy threshold. To ensure the model accuracy of FedKD, we choose to set a higher as the model training continues to learn more accurate models.

5 Conclusion

In this paper, we propose a communication efficient federated learning method based on knowledge distillation named FedKD. In our approach, we propose an adaptive mutual distillation method to reciprocally learn a teacher model and a student model on each client, where the distillation intensity is controlled by their prediction correctness. The large teacher model is locally updated, while the small student model is shared among different clients and learned collaboratively, which can effectively reduce the communication cost. In addition, we propose a dynamic gradient approximation method to further reduce the communication cost. Extensive experiments on two benchmark datasets for different tasks validate that FedKD can largely reduce the communication cost in federated learning while keeping promising model performance.

References

- Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long-and short-term user representations. In *ACL*, pages 336–345.
- Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. 2018. Large scale distributed neural network training through online distillation. In *ICLR*.
- Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudomasked language models for unified language model pre-training. In *ICML*, pages 642–652. PMLR.
- Yoshua Bengio and Yann LeCun. 2015. Adam: A method for stochastic optimization. In *ICLR*.
- Sebastian Caldas, Jakub Konečny, H Brendan McMahan, and Ameet Talwalkar. 2018. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT, pages 4171–4186.
- Suyu Ge, Fangzhao Wu, Chuhan Wu, Tao Qi, Yongfeng Huang, and Xing Xie. 2020. Fedner: Medical named entity recognition with federated learning. *arXiv* preprint arXiv:2003.09288.
- Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2018. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv* preprint arXiv:1503.02531.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Trung Huynh, Yulan He, Alistair Willis, and Stefan Rueger. 2016. Adverse drug reaction classification with deep neural networks. In *COLING*, pages 877–887
- Di Jiang, Yuanfeng Song, Yongxin Tong, Xueyang Wu, Weiwei Zhao, Qian Xu, and Qiang Yang. 2019. Federated topic modeling. In *CIKM*, pages 1071–1080.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling BERT for natural language understanding. In *EMNLP Findings*, pages 4163–4174.

- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746– 1751.
- Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.
- Daliang Li and Junpu Wang. 2019. Fedmd: Heterogenous federated learning via model distillation. *arXiv* preprint arXiv:1910.03581.
- Guanyu Lin, Feng Liang, Weike Pan, and Zhong Ming. 2020a. Fedrec: Federated recommendation with explicit feedback. *IEEE Intelligent Systems*.
- Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. 2020b. Ensemble distillation for robust model fusion in federated learning. *arXiv preprint arXiv*:2006.07242.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pages 1273–1282.
- Zulfat Miftahutdinov, Ilseyar Alimova, and Elena Tutubalina. 2019. Kfu nlp team at smm4h 2019 tasks: Want to extract adverse drugs reactions from tweets? bert to the rescue. In *SMM4H*, pages 52–57.
- Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *KDD*, pages 1933–1942. ACM.
- Tao Qi, Fangzhao Wu, Chuhan Wu, Yongfeng Huang, and Xing Xie. 2020. Privacy-preserving news recommendation model learning. In *EMNLP: Findings*, pages 1423–1432.
- Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, and Raman Arora. 2020. Fetchsgd: Communication-efficient federated learning with sketching. In *ICML*, pages 8253–8265. PMLR.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv* preprint arXiv:1910.01108.
- Hyowoon Seo, Jihong Park, Seungeun Oh, Mehdi Bennis, and Seong-Lyun Kim. 2020. Federated knowledge distillation. *arXiv preprint arXiv:2011.02367*.
- Reza Shokri and Vitaly Shmatikov. 2015. Privacypreserving deep learning. In CCS, pages 1310– 1321.

- Dianbo Sui, Yubo Chen, Jun Zhao, Yantao Jia, Yuantao Xie, and Weijian Sun. 2020. Feded: Federated learning via ensemble distillation for medical relation extraction. In *EMNLP*, pages 2118–2128.
- Lichao Sun and Lingjuan Lyu. 2020. Federated model distillation with noise-free differential privacy. *arXiv preprint arXiv:2009.05537*.
- Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. In *EMNLP-IJCNLP*, pages 4314–4323.
- Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*.
- Heyuan Wang, Fangzhao Wu, Zheng Liu, and Xing Xie. 2020a. Fine-grained interest matching for neural news recommendation. In *ACL*, pages 836–845.
- Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Dkn: Deep knowledge-aware network for news recommendation. In *WWW*, pages 1835–1844.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020b. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*.
- Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez. 2018. Overview of the third social media mining for health (smm4h) shared tasks at emnlp 2018. In *SMM4H*, pages 13–16.
- Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019a. Neural news recommendation with attentive multiview learning. In *IJCAI*, pages 3863–3869.
- Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019b. Npa: Neural news recommendation with personalized attention. In *KDD*, pages 2576–2584.
- Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019c. Neural news recommendation with multi-head self-attention. In *EMNLP*, pages 6390–6395.
- Chuhan Wu, Fangzhao Wu, Zhigang Yuan, Junxin Liu, Yongfeng Huang, and Xing Xie. 2019d. Msa: Jointly detecting drug name and adverse drug reaction mentioning tweets with multi-head self-attention. In *WSDM*, pages 33–41. ACM.
- Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *ACL*, pages 3597–3606.
- Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *TIST*, 10(2):1–19.

- Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *CVPR*, pages 4320–4328.
- Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Attention-based lstm network for cross-lingual sentiment classification. In *EMNLP*, pages 247–256.
- Ligeng Zhu and Song Han. 2020. Deep leakage from gradients. In *Federated Learning*, pages 17–31. Springer.

A Appendix

A.1 Comparison with Additional Baselines

To provide benchmarks on the MIND and SMM4H datasets, we compare the performance of our FedKD approach with several baseline methods on these datasets. On the MIND dataset, the additional baseline methods to be compared include: (1) EBNR (Okura et al., 2017), embedding-based news recommendation with GRU; (2) DKN (Wang et al., 2018), deep knowledge network for news recommendation; (3) NPA (Wu et al., 2019b), news recommendation with personalized attention; (4) NAML (Wu et al., 2019a), news recommendation with attentive multi-view learning; (5) LSTUR (An et al., 2019), news recommendation with long short-term user interest; (6) NRMS (Wu et al., 2019c), news recommendation with multi-head self-attention; (7) FIM (Wang et al., 2020a), finegrained interest matching for news recommendation. On the SMM4H dataset, we compare with the following baseline methods: (1) CNN (Kim, 2014), CNN for text classification; LSTM (Hochreiter and Schmidhuber, 1997), long short-term memory network; (3) CNN+Att (Huynh et al., 2016), using attentive pooling after CNN model; (4) LSTM+Att (Zhou et al., 2016), applying attention pooling to LSTM; (5) MSA (Wu et al., 2019d), a multi-head self-attention based approach for ADR detection; (6) BERT (Miftahutdinov et al., 2019), using BERT for ADR detection. The results on the MIND and SMM4H datasets are respectively shown in Tables 4 and 5.¹⁰ From the results, we find the performance of our FedKD approach consistently outperform all the baseline methods (e.g., 70.7% v.s. 68.5% AUC scores on MIND). The further t-test results also show the differences between FedKD and other baseline methods are significant (p < 0.05). This is because our approach takes the advantage of the state-of-the-art pre-trained language models and allows the teacher and student models to collaboratively learn from each other, which are helpful for learning strong models.

A.2 Additional Results on SMM4H

We also report the additional results on the *SMM4H* dataset, which are shown in Figs. 8-11. We observe similar phenomena with the results on *MIND*.

Methods	AUC	MRR	nDCG@5	nDCG@10
EBNR	66.1±0.3	31.9±0.3	34.9 ± 0.3	40.5±0.4
DKN	65.2 ± 0.3	31.5 ± 0.3	34.1 ± 0.2	39.8 ± 0.3
NPA	67.4 ± 0.2	32.6 ± 0.3	35.5 ± 0.3	41.3 ± 0.3
NAML	67.4 ± 0.2	32.5 ± 0.2	35.4 ± 0.2	41.2 ± 0.2
LSTUR	67.9 ± 0.3	33.0 ± 0.3	35.9 ± 0.3	41.8 ± 0.3
NRMS	$68.2 {\pm} 0.2$	33.4 ± 0.2	36.3 ± 0.1	42.1 ± 0.2
FIM	68.5 ± 0.3	33.6 ± 0.2	36.6 ± 0.3	42.4 ± 0.3
FedKD ₄	71.0 ±0.1	35.6 ±0.1	38.9 ±0.1	44.8 ±0.1
$FedKD_2$	70.5 ± 0.1	35.3 ± 0.2	38.6 ± 0.1	44.3 ± 0.2

Table 4: Performance of different methods on MIND.

Methods	Precision	Recall	Fscore
CNN	48.3 ± 1.0	52.1±1.1	50.2±0.6
LSTM	49.6 ± 1.2	50.5 ± 0.8	50.0 ± 0.7
CNN+Att	48.3 ± 0.9	53.0 ± 0.9	50.5 ± 0.5
LSTM+Att	$48.5 {\pm} 0.5$	52.7 ± 0.7	50.5 ± 0.4
MSA	51.2 ± 0.6	53.8 ± 0.8	52.4 ± 0.4
BERT	56.6 ± 0.6	59.8 ± 0.9	58.0 ± 0.6
FedKD ₄	59.4 ±0.6	62.8 ±0.9	60.7 ±0.5
$FedKD_2$	58.2 ± 0.7	62.4 ± 0.9	$59.8 {\pm} 0.6$

Table 5: Performance of different methods on SMM4H.

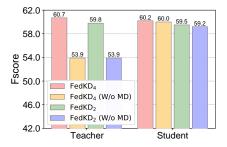


Figure 8: Influence of mutual distillation on the student and teacher models.

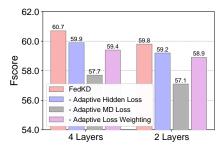


Figure 9: Effect of adaptive mutual distillation.

A.3 Algorithm Workflow

The workflow of *FedKD* is summarized in Algorithm 1.

A.4 Experimental Environment

Our experimental environment is built on a Linux server with Ubuntu 16.04 operation system. The version of Python is 3.6 The server has 4 Tesla

¹⁰The baselines are trained on a centralized data storage.

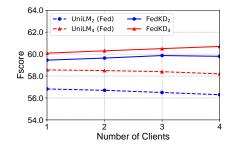


Figure 10: Influence of client number.

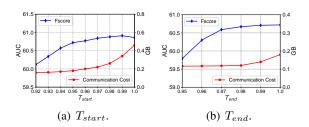


Figure 11: Influence of T_{start} and T_{end} on model performance and communication cost.

V100 GPUs with 32GB memory. The CPU type is Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz. The total memory is 128GB. We use the horovod framework for parallel model training on the 4 GPUs, each of which represents a platform.

A.5 Model Initialization

In our approach, we use the token embedding layer and the first 4 or 2 layers of UniLM to initialize the student model. We do not change the hidden dimension of the model because the UniLMv2 models with other hidden dimensions are not released. Note that our approach does not have limitations on the hidden dimension of the student model.

A.6 Running Time

On the MIND dataset, the total training of time $FedKD_4$ and $FedKD_2$ are take around 66 and 57 hours, respectively. On the SMM4H dataset, their total training times are about 12 minutes and 10.5 minutes, respectively.

A.7 Hyperparameter Settings

The complete hyperparameter settings are listed in Table 6. The negative sampling ratio means the number of negative samples packed with each positive sample for model training. We use the crossentropy loss to classify which sample is the positive sample. The over sampling ratio means the repeating number of positive samples due to the high class imbalance.

Algorithm 1 FedKD

- 1: Setting the teacher learning rate η_t and student learning rate η_s , client number N
- 2: Setting the hyperparameters T_{start} and T_{end}
- 3: for each client i (in parallel) do
- 4: Initialize parameters Θ_i^t , Θ^s
- 5: repeat

6:

15:

- $\mathbf{g}_{i}^{t},\mathbf{g}_{i}$ =LocalGradients(i)
- 7: $\Theta_i^t \leftarrow \Theta_i^t \eta_t \mathbf{g}_i^t$
- 8: $\mathbf{g}_i \leftarrow \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i$
- 9: Clients encrypt $\mathbf{U}_i, \mathbf{\Sigma}_i, \mathbf{V}_i$
- 10: Clients upload U_i, Σ_i, V_i to the server
- 11: Server decrypts U_i, Σ_i, V_i
- 12: Server reconstructs \mathbf{g}_i
- 13: Global gradients $\mathbf{g} \leftarrow 0$
- 14: **for** each client i (in parallel) **do**
 - $\mathbf{g} = \mathbf{g} + \mathbf{g}_i$
- 16: **end for**
- 17: $\mathbf{g} \leftarrow \mathbf{U} \mathbf{\Sigma} \mathbf{V}$
- 18: Server encrypts U, Σ, V
- 19: Server distributes U, Σ, V to user clients
- 20: Clients decrypt U, Σ, V
- 21: Clients reconstructs g
- 22: $\Theta^s \leftarrow \Theta^s \eta_s \mathbf{g}/N$
- 23: **until** Local models converges
- 24: end for

LocalGradients(*i*):

- 25: Compute task losses $\mathcal{L}_{t,i}^t$ and $\mathcal{L}_{s,i}^t$
- 26: Compute losses $\mathcal{L}_{t,i}^d$, $\mathcal{L}_{s,i}^d$, $\mathcal{L}_{t,i}^h$, and $\mathcal{L}_{s,i}^h$
- 27: $\mathcal{L}_{i}^{t} \leftarrow \mathcal{L}_{t,i}^{t} + \mathcal{L}_{t,i}^{d} + \mathcal{L}_{t,i}^{h}$
- 28: $\mathcal{L}_{i}^{s} \leftarrow \mathcal{L}_{s,i}^{t} + \mathcal{L}_{s,i}^{d} + \mathcal{L}_{s,i}^{h}$
- 29: Compute local teacher gradients \mathbf{g}_i^t from \mathcal{L}_i^t
- 30: Compute local student gradients \mathbf{g}_i from \mathcal{L}_i^s
- 31: **return** $\mathbf{g}_i^t, \mathbf{g}_i$

Urmannamatana	MIND	SMM4H
Hyperparameters	MIND	ЗИИ4П
LM hidden dimension	768	768
CNN feature map dimension	256	256
LSTM hidden dimension	256	256
negative sampling ratio	4	-
over sampling ratio	-	2 for LMs, 9 for others
attention query dimension	200	200
dropout	0.2	0.2
optimizer	Adam	Adam
teacher model learning rate	2e-6	2e-6
student model learning rate	5e-6	1e-5
batch size	32	64
T_{start}	0.95	0.95
T_{end}	0.98	0.98
Epoch	3	2

Table 6: Hyperparameter settings.