Cooperative Learning via Federated Distillation over Fading Channels

*Jinhyun Ahn, **Osvaldo Simeone, and *Joonhyuk Kang
*KAIST, South Korea,
**King’s College London, UK
Contents

- Introduction
- Problem Definition
- Proposed Method
- Numerical Results
- Conclusion
- References
- Appendix
Introduction

- Federated Learning
  - Federated learning (FL), is developed recently, which features distributed learning at edge devices and periodic local-update of model (model coefficients or gradients) averaging at an parameter server (PS)
  - Nevertheless, the updates uploading in FL can be still bandwidth-consuming as an AI model usually comprises millions to billions of parameters [7]
  - A key research issue that is particularly hot recently is to reduce the overhead in update uploading to further accelerate the model training process [8]–[13]
    - Addressing the straggler effect in synchronous update averaging
    - Developing lazily updating algorithm that schedules only those devices with significant updates to save the updating bandwidth
    - Compress gradient vectors by exploiting its inherent sparsity (most of the gradient elements are insignificant and thus can be truncated without harming the model accuracy)
Introduction

- Federated distillation
  - To alleviate this problem, federated distillation (FD) was introduced for classification problems in [14]
  - Distillation for learning model was proposed by Hinton et al. [15]
    - To transfer a knowledge about a learning model, output vectors per inputs are sent from teacher
      \[
      \begin{pmatrix}
      1 \\
      0 \\
      0 \\
      0 \\
      0 \\
      0 \\
      0 \\
      0
      \end{pmatrix}
      \]
      \[
      \begin{pmatrix}
      0.611 \\
      0.032 \\
      0.002 \\
      0.015 \\
      0.012 \\
      0.005 \\
      0.006 \\
      0.231 \\
      0.075 \\
      0.011
      \end{pmatrix}
      \]
      \[
      \begin{pmatrix}
      0.805 \\
      0.016 \\
      0.001 \\
      0.008 \\
      0.006 \\
      0.002 \\
      0.003 \\
      0.116 \\
      0.038 \\
      0.005
      \end{pmatrix}
      \]
    - Cross entropy with Per input
    - Weighted average

- In FD, devices periodically exchange the average output logit vectors per labels instead of local update of model in FL (less information but lower accuracy gain than FL)
  - We propose a novel hybrid federated distillation (HFD) scheme that aims at bridging the performance gap between FD and FL
Introduction

- **Wireless Implementation of FD and HFD**
  - In many practical implementations, however, bandwidth of the communication channel from devices to the PS turns out to be the main bottleneck [16], [17]
  - Recently, a multiple access scheme called “over-the-air” computation (AirComp) is particularly appealing in the scenario as it integrates transmission and computation and allows “one-shot” data aggregation by exploiting the waveform-superposition property of a multi-access channel (MAC) [18-19], [25]
  - There is no previous work about the wireless implementation of FD
  - We propose a communication scheme for the implementation of FD focusing on the quantization and compression
  - Both a conventional digital scheme and an analog scheme are considered for the communication in the uplink and downlink
Problem Definition

- Problem Definition
  - $K$ devices communicate via an Access Point (AP) so as to train a machine learning model that outperforms a model trained solely on the local training set.
  - For device $k$:
    - Data set $\mathcal{D}_k : (c, t)$ (vector of covariates, one-hot encoding vector)
    - Trains its own neural network model: $\mathbf{w}_k, \ W \times 1$
    - Neural network produces the logit vector: $\mathbf{s}(\mathbf{c} | \mathbf{w}_k)$
      - the probability vector: $\hat{t}(c | \mathbf{w}_k)$

$$
\hat{t}(c | \mathbf{w}_k) = \frac{1}{L} \sum_{i=1}^{L} e^{s_i} = \begin{bmatrix} e^{s_1} \\ \vdots \\ e^{s_L} \end{bmatrix}
$$
Training Protocols (FL)

**Algorithm 2** Federated Learning (FL)

for each iteration $i = 1, \ldots, I$

for each device $k = 1, \ldots, K$

**download** from PS the average weight update

$$\Delta w_{i-1} = \frac{1}{K} \sum_{k=1}^{K} \Delta w_{i-1}^k$$

set initial value

$$w_i^k = w_{i-1}^k + \Delta w_{i-1} - \Delta w_{i-1}^k \equiv w_{i,o}^k$$

for each iteration of local training

**do** SGD update as in (1), for a randomly selected training example $(c, t) \in D_k$

do

**upload** update $\Delta w_i^k = w_i^k - w_{i,o}^k$ to PS

- The weight vectors at each device are initialized to the average weight vectors using the average weight update downloaded from the PS
- Devices carry out a number of local updates using SGD as the update in IL
- Upload the resulting weight vector to the PS
Training Protocols (FD)

- Instead of neural network model parameters, devices exchange the local-averaged logit vector per labels (10 values per 10 classes for MNIST).
- Applies the global-averaged logit vectors per labels for its own local training.
Training Protocols (FD)

**Algorithm 3 Federated Distillation (FD)**

for each iteration $i = 1, \ldots, I$

for each device $k = 1, \ldots, K$

download from PS the global-averaged logit vectors for all labels $t = 1, \ldots, L$

$$s_{i,t} = \frac{1}{K} \sum_{k'=1}^{K} s_{i,t}^{k'}$$ (2)

obtain the local logit vectors

$$s_{i,t}^{k} = \frac{Ks_{i,t} - s_{i,t}^{k}}{K - 1}$$ (3)

initialize $s_{i+1,t}^{k} \leftarrow 0$ and $n_{i+1,t}^{k} \leftarrow 0$ for all labels $t = 1, \ldots, L$

for each iteration of local training

do SGD update

$$w_{i}^{k} \leftarrow w_{i}^{k} - \alpha \nabla w_{i}^{k} \left\{ (1 - \beta) \phi \left( i \left( c | w_{i}^{k} \right), t \right) + \beta \phi \left( i \left( c | w_{i}^{k} \right), \hat{i} (s_{i,t}^{k}) \right) \right\}$$ (4)

for a randomly selected training example $(c, t) \in D_{k}$

update the logit vector and the label counter

$$s_{i+1,t}^{k} \leftarrow s_{i+1,t}^{k} + s \left( c | w_{i}^{k} \right)$$

$$n_{i+1,t}^{k} \leftarrow n_{i+1,t}^{k} + 1$$

end

upload the local-averaged logit vectors $s_{i+1,t}^{k} \leftarrow s_{i+1,t}^{k}/n_{i+1,t}^{k}$ to the PS for all labels $t = 1, \ldots, L$

- In (2) and (3), each device excludes its own information from the averaged logit vectors
- In (4), each device carries out a number of local updates using the averaged logit vectors as a regularizer
- During the local updates, each device computes and uploads the local-averaged logit vectors for all labels to the PS
Training Protocols (HFD)

- The proposed HFD modifies FD by using not only the average logit vector but also the average covariate vector per label, which is shared during a preliminary offline phase.
- In the distillation [15],
  - Teacher and students share the same covariates vectors
  - The teacher’s knowledge is transferred by sending every logit vectors for all covariates
  - Student uses associated logit vectors for local training of covariates
- In FD, the teacher’s knowledge is the average logit vectors per labels
- In HFD,
  - The teacher’s knowledge is average covariate vector and its output logit vectors per labels
  - Updates consist of distillation phase and IL phase
    - **Distillation phase**: updates over only global averaged covariate vectors using the downloaded logit vectors as regularizer as in FD
    - **IL phase**: updates over local dataset
Prior to the global iterations
- Obtain the local averaged covariate vectors
  \[ \tilde{c}_t^k \quad t = 1, \ldots, L \]
- Download the global averaged covariate vectors exclude its own information
  \[ \tilde{c}_t = \frac{1}{K} \sum_{k' = 1}^{K} \tilde{c}_t^{k'} \]
  \[ \tilde{c}_t^k = \frac{K \tilde{c}_t - \tilde{c}_t^k}{K - 1} \]

Algorithm 4 Hybrid Federated Distillation (HFD)

for each device \( k = 1, \ldots, K \)
  for each iteration \( i = 1, \ldots, I \)
    - download from PS the global-averaged logit vectors (5) for all labels \( t = 1, \ldots, L \)
    - obtain the logit vectors (6)
    - for each iteration of the distillation phase of local training
      do SGD update as in (7) for a data point \( (\tilde{c}_t^k, t) \) for a randomly chosen label \( t \)
    - for each iteration of the IL phase of local training
      do SGD update as in (3) for a randomly selected training example \( (c, t) \in D \)
  - upload the logit vectors
    \[ s_{i+1,t}^k = s(\tilde{c}_t^k \mid w_i^k) \]
  - to the PS for all labels \( t = 1, \ldots, L \)

- As in FD, each device downloads the global averaged logit vectors (and exclude)
- At the distillation phase, does SGD updates with the covariate vectors using the logit vectors as a regularizer for a randomly chosen label
- At the IL phase, each device does SGD updates with its own local dataset
- After the local updates, computes and uploads the output logit vectors of local averaged covariate vectors per labels
Wireless Cooperative Training

- Proposed four wireless implementations of FL and FD/HFD
  - Digital (D) or analog (A) communication in uplink and downlink
  - digital-digital (D-D) / digital-analog (D-A)
  - analog-digital (A-D) / analog-analog (A-A)

- Digital transmission for both uplink and downlink is based on separate source-channel coding
  - UL: Equal resource allocation to devices, sparsification and quantization(FD/HFD)
  - DL: Broadcast after compression and quantization

- Analog transmission implements joint source-channel coding through over-the-air computing
  - UL: Simultaneous transmission in uncoded manner
  - DL: Broadcast $\rightarrow$ Consider scaling factor and AMP algorithm at each device
Wireless Cooperative Training

- **Channel Model**

  - During each information exchange phase of the \(i\)-th global iteration, devices share a **fading uplink multiple-access channel**: The received signal is
    \[
    y_i = \sum_{k=1}^{K} h_i^k x_i^k + z_i
    \]
    - \(h_i^k\): quasi-static fading channel from the device \(k\) to the AP
    - \(x_i^k\): \(T_U \times 1\) signal transmitted by the device \(k\)
    - \(z_i\): \(T_U \times 1\) noise vector with i.i.d. \(\mathcal{CN}(0,1)\) entries
    - Each device \(k\) has a power constraint \(E[\|x_i^k\|_2^2]/T_U \leq P_U\)

  - The AP can **broadcast to all device in downlink** so that the received signal is
    \[
    y_i^k = g_i^k x_i + z_i^k
    \]
    - \(g_i^k\): quasi-static fading channel from the AP to the device \(k\)
    - \(x_i\): \(T_D \times 1\) signal transmitted by the AP
    - \(z_i^k\): \(T_D \times 1\) noise vector with i.i.d. \(\mathcal{CN}(0,1)\) entries
    - The AP has a power constraint \(E[\|x_i\|_2^2]/T_D \leq P_D\)
Wireless Cooperative Training

- **Performance Comparison**
  - 10 devices train a 6-layer CNN to carry out image classification based on subsets of the MNIST data set available at each device
  - The distributions of dataset are i.i.d.
    - Randomly select disjoint sets of 64 samples from the 60,000 training MNIST examples, and allocate each set to a device
  - Channel fading: Rician fading
  - Number of global iteration: 10
  - Learning rate: 0.001
  - Number of quantization bits: 16
  - Sparsification level for analog transmission: \(q = 4T/5\)
  - \(T_U = T_D = T\)
  - \(P_D = P_U + 10\) dB
Wireless Cooperative Training

- **Performance Comparison**

  - Number of channel uses varies under $P_U = 0 \text{ dB}$
  - FD and HFD significantly outperform FL at low values of $T$ that is, with limited spectral resources
  - HFD is seen to uniformly improve over FD
  - The A-A scheme is clearly preferable over the alternatives

![Graph](image)

*Fig. 2: Classification test accuracy for IL, FL, FD, and HFD under implementations D-D, D-A, A-D, and A-A*
Wireless Cooperative Training

- **Performance Comparison**

  - The number $T$ is 2500
  - The figure confirms that FD and HFD significantly outperform FL at low values of $P$.
  - And HFD uniformly improves over FD.
  - The A-A scheme shows the best performance, especially for lower values of $P$.
  - It is checked that the performance of analog transmission scheme converges when $P$ increases (The figure should be plotted for larger SNR).

Fig. 2: Classification test accuracy for IL, FL, FD, and HFD under implementations D-D, D-A, A-D, and A-A.
Development of FD/HFD to support FL under limited communication resources
- Propose the HFD training protocol
- Investigate the wireless implementations of FD/HFD

Questions → wlsgus3396@kaist.ac.kr
References


References

Appendix: Wireless Cooperative Training

- **Uplink Digital Transmission (FL, FD/HFD)**
  - Consider for simplicity an equal resource allocation to devices
  - The number of bits that can be transmitted from each device $k$ at the $i$-th global iteration is given using Shannon’s capacity
    \[ B_{U,k,i} = \frac{T_U}{K} \log_2 \left( 1 + |h_i^k|^2 K P_U \right) \]
  - Each device $k$ compresses the corresponding information to be sent to the AP to no more than $B_{U,k,i}$ bits
  - Devices are aware of the rate and hence of the channel power
  - AP has full channel state information
Appendix: Wireless Cooperative Training

- **Uplink Digital Transmission (FL)**
  - Each device $k$ aims to send $\Delta w_i^k$ at the $i$-th global iteration
  - Adopts spares binary compression with error accumulation as
    \[ v_i^k = \text{sparse}_{q_i^k} (\Delta w_i^k + \Delta_i^k) \]
    where the accumulated quantization error is updated as
    \[ \Delta_i^{k+1} = \Delta w_i^k + \Delta_i^k - Q_b (v_i^k) \]
  - Then it sends
    \[ B_{U,k,i}^{FL} = b + \log_2 \left( \frac{W}{q_i^k} \right) \]
    bits to send the value $Q_b (\mu)$ and the indices of the non-zero elements of $v_i^k$, where $q_i^k$ is chosen as the largest integer satisfying $B_{U,k,i}^{FL} \leq B_{U,k,i}$

\[ \begin{align*}
\text{sparse}_q (u) & \quad - \text{All elements except the largest } q \text{ elements and smallest } q \text{ elements of } u \text{ are set to zero} \\
\mu^+ & \quad - \text{mean of remaining positive elements} \\
\mu^- & \quad - \text{mean of remaining negative elements} \\
\text{If } |\mu^-| > |\mu^+| & \quad - \text{the negative elements are set to zero and the positive elements are set to } \mu^+ \\
\text{If } |\mu^-| < |\mu^+| & \quad - \text{the positive elements are set to zero and the negative elements are set to } \mu^- \\
\end{align*} \]

- Quantizes each non-zero element of $u$ using a uniform quantizer with $b$ bits per each non-zero element
Appendix: Wireless Cooperative Training

- Uplink Digital Transmission (FD/HFD)
  - Each device $k$ aims to send logit vectors $s_{i,t}^k$ at the $i$-th global iteration for all labels $t = 1, \ldots, L$
  - Adopts sparsification and quantization as
    \[
    q_{i,t}^k = Q_b(\text{thresh}_{q_i^k}(s_{i,t}^k)) \quad t = 1, \ldots, L
    \]
  - Then it sends
    \[
    B_{U,k,i}^{FD} = L(bq_i^k + \log_2(\frac{L}{q_i^k}))
    \]
    bits to send the non-zero values and the indices of the non-zero elements of $q_{i,t}^k$
    where $q_i^k$ is chosen as the largest integer satisfying $B_{U,k,i}^{FD} \leq B_{U,k,i}$

\[\text{thresh}_{q}(u)\]

- Sets all elements of the input vector $u$ to zero except the $q$ elements with the largest absolute values
Appendix: Wireless Cooperative Training

- **Downlink Digital Transmission (FL, FD/HFD)**
  - **The number of bits** that can be transmitted from AP to devices at the $i$-th global iteration is given using Shannon’s capacity
    
    $B_{D,i} = \min_k \left( T_D \log_2 \left( 1 + |g_i^k|^2 P_D \right) \right)$

  - **Satisfying** $B_{D,i}^{FL} \leq B_{D,i}$ and $B_{D,i}^{FD} \leq B_{D,i}$, AP compresses and quantizes the corresponding information
Appendix: Wireless Cooperative Training

- Uplink Analog Transmission (FL, FD/HFD)
  - All the devices transmit their information simultaneously in an uncoded manner to the AP
  - Different types of power control at each devices have been studied in the literature, namely full-power transmission, channel inversion [18],[19], and optimized power control [26], [27]
  - In this paper, full-power transmission is considered for simplicity
  - Each device have knowledge of the phase of the channel to the AP, and the AP has full channel state information
  - In analog transmission of a vector, only the values of number of channel uses can be sent (usually much less than the number of network model coefficients)
  - The gradient update should be sparsified and compressed into a smaller dimension
  - The PS recovers the sum of gradient updates by applying AMP (approximate message passing)
  - It is assumed that the gradient updates have similar sparsity pattern among the devices under the i.i.d. data distribution
Appendix: Wireless Cooperative Training

- Uplink Analog transmission (FL)
  - Each device $k$ aims to send $\Delta w_i^k$ at the $i$-th global iteration
  - In order to enable dimensionality reduction, a pseudo-random matrix $A_U \in \mathbb{R}^{2TU \times W}$ with i.i.d. entries $\mathcal{N}(0, 1/2T_U)$ is generated and shared
  - Each device $k$ computes and $v_i^k = \text{thresh}_q \left( \Delta w_i^k + \Delta_i^k \right)$ for sparsification
  - To transmit dimension reduced vector $\hat{v}_i^k = A_U v_i^k$, transmit $x_i^k \in \mathbb{C}^{TU \times 1}$,
    \[
    x_i^k (m) = \hat{v}_i^k (2m - 1) + j \hat{v}_i^k (2m), m = 1, \ldots, TU
    \]
  - Each device $k$ transmits $\gamma_i^k e^{-j h_i^k} x_i^k \in \mathbb{C}^{TU \times 1}$, $\gamma_i^k = \sqrt{P_U T_U / \|x_i^k\|_2}$ for full power transmission
  - The PS scales the received signal by
    \[
    \nu_i = \frac{\sum_{k'=1}^{K} \gamma_{i}^{k'} |h_i^{k'}|}{\frac{1}{2} + \sum_{k'=1}^{K} (\gamma_{i}^{k'} |h_i^{k'}|)^2}
    \]
    for minimum mean square error estimate of the sum $A_U \sum_{k=1}^{K} v_i^k$
  - The PS applies AMP algorithm to recover $\sum_{k=1}^{K} v_i^k$
Appendix: Wireless Cooperative Training

- **Uplink Analog transmission (FD)**
  - Each device $k$ aims to send $s_{i,t}^k$ at the $i$-th global iteration $t = 1, \ldots, L$
  - Apply repetition coding since $L^2$ is usually lower than $2TU$
  - Each device applies repetition coding with the source integer bandwidth expansion factor $\rho = \lceil 2TU/L^2 \rceil \geq 1$
  - And compute
    \[
    \mathbf{v}_i^k = \mathbf{R}_\rho \mathbf{s}_i^k \in \mathbb{R}^{\rho L^2 \times 1}
    \]
    \[
    \mathbf{R}_\rho = 1_\rho \otimes \mathbf{I}_{L^2}
    \]
    \[
    1_\rho = (1, \ldots, 1)^T
    \]
    \[
    \mathbf{s}_i^k = [(s_{i,1}^k)^T, \ldots, (s_{i,L}^k)^T]^T
    \]
  - And transmit as the same way with case of FL
  - AP multiplies $\mathbf{R}_\rho^T / \rho$ to estimate $\sum_{k=1}^K \mathbf{v}_i^k$
Appendix: Wireless Cooperative Training

- Downlink Analog Transmission (FL, FD/HFD)
  - For the downlink broadcast communication from AP to devices,
    - The AP transmits with full power in a same manner of each device at the uplink
    - Each device applies a scaling factor and the AMP algorithm in order to estimate the vector transmitted by the AP, in a similar manner of AP at the uplink