Intro

안녕하세요 이현주입니다. 제가 선택한 논문은 고려대학교 차세대통신 백상헌 교수님이 쓰신 FL을 위한 상황인지와 양자화 레벨 세션입니다.

Hi, I’m hyeonju lee and today I’m gonna talk about situations-aware cluster and Quantization level section algorithm for fast federated learning, which is written by Professor Sangheon Beak, who is in the field of next-generation communications at Korea university.

To put it simply, apply SITUA-CQ algorithm for federated learning.

**FL이란?**

And What is federated learning?

This is distributed learning to Contrary concept federated learning. In this process, gather clients data to centralized server. And In server, train model using client’s data and share the results of model. When exchanging data in Severs, privacy issues can arise. For example, if in a hospital, we want train patients’ data, their data go out to external networks and it makes privacy issue.

Federated learning is the process, not sharing their own data. that means FL can preserve privacy of the device. It’s like doctor’s conference with their experience of individual surgery patients’ data

Note that each client may have its own data distribution and it’s problem of federated learning

반대 개념인 분산학습은 user들의 데이터를 centralized server에 모으고 서버에 있는 model을 학습시켜 유저들에게 결과를 전달하는 방식이다. 여기서 문제는 privacy가 담긴 data를 주고받을 때 발생한다. 예를들어 병원의 환자를 user라고 보고 환자의 data를 가지고와 학습시키려 할 때, 환자의 데이터를 외부망에 보내야하는데 이때 개인정보보호 문제가 발생한다.

FL은 최신의 AI 모델을 다운받는다. 그 후 각자가 가지고 있는 데이터를 기반으로 모델을 학습시킨 후 그 모델을 클라우드로 올려보낸다. 클라우드는 이러한 모델 값을 기반으로 새로운 모델을 업데이트 한다. 다시 비유하자면 각 의사들이 각 병원에서 환자를 치료한 후 한 군데 모여서 각자의 수술 경험에 대해 이야기하는것이다.

이렇게 데이터가 이동하지 않기 때문에 파워 소비와 latency에서 이득이다. 프라이버시 문제 해결또한

FL 모델의 문제점은 클라이언트가 가지고 있는 데이터가 동일한 확률 분포를 가지고 있다는 점이다.

**시스템 과정**

Let’s move on system process. And this process consists of six stages:

These stages are repeated until target accuracy is achieved.

Step1: performance reporting

Each client report it’s data distribution, computational power, rate of data uplink and downlink to FL server

Step2: Cluster/ Step3: round time estimation

Next, in the clustering stage, the FL server estimates the individual round time of each client.

Step4: Quantization level section

. Based on the expected cluster round times, the FL server allocates the quantization levels for the chosen clusters

**+what is Quantization level**?

In network communication, we get continuous analog signal.

We need to convert a continuous analog signal to digital signal. And setting the data size is Quantizaiton. Increasing the quantization level gives more detailed representation.

Step5: local training

I’d already told you that for FL clients train deep model using their own local data out of server.

the deep model is sent to the FL server.

Step6: Aggregation

the final aggregation stage,, it aggregates all deep models for the next round by averaging parameters of deep models.

데이터 분산을 다양하게 가진 클라이언트를-> 클러스터 하는 것

선택된 클러스터에서 양자화 수준을 할당하는 것

6단계의 과정으로 이루어져있다.

1단계 – 수행보고

각 클라이언트들은 데이터분산, 컴퓨터 파워, 업링크/다운링크 데이터 비율 측정 후 서버에 보고한다. 클라이언트 마다 다른 값을 가지고 있기 때문에

2단계 – 클러스터링/3단계 – 클러스터 라운드타임 측정

FL서버는 각 클라이언트의 개인 라운드 시간 추정한다. 그것을 기반으로 클라이언트를 그룹화.

4단계 – 양자화 수준 선택

예상된 클럿스터 라운드 시간을 기반으로 FL서버는 클러스터 선택 후 양자화 레벨 할당

!!양자화 단계란 무엇인가

연속적인 신호 값을 디지털 신호로 변환하는 과정에서 데이터의 크기를 설정하는 것이다

양자화 레벨이 커질수록 세밀한 표현이 가능하다

5단계 – 로컬훈련

FL서버가 각 클러스터의 클라이언트에게 딥모델을 보내고 클라이언트는 로컬 데이터로 딥 모델을 학습시킴

6단계 – 집합

FL서버가 클러스터로부터 딥 모델을 모두 받고 그것의 파라미터를 평균화 하여 딥 모델 업데이트,

**Deep model 구조**

The deep model consists of six convolutional layers and two fully connected layers. The six convolutional layers have two sets of 32, 64, 128 each, respectively, and ReLu and the max-pooling layers which filter size is 2 × 2 are followed.

, there are 10 classes, and it consists of 50,000 training images and 10,000 test images

6개의 layer, 32,64,126개의 필터로 이루어져있다. Relu, max-pooling 함수 사용

**정확도**

I’d like to mention three points to effect accuracy, Data distribution, number of clusters, quantization levels.

1. 전체적인 클러스터의 데이터 분산의 정확도의 영향

DkL(kull-back diversence)의 감소함에 따라 정확도는 급격하게 증가하는것을 볼 수 있다.

이 결과는 FL서버는 각 **클러스터의 충분히 짧은 분산 거리를 적용해** 클라이언트를 클러스터로 그룹해야만한다.

the range of DKL (kull-back diversence) is distance of distribution data.

It can be found that the accuracy increases rapidly with the decrease of DKL..

If you look at this table you can see that ., green, yellow, red ones are small DkL and high accuracy.

This result indicates that the FL server should group the clients into clusters to attain sufficiently short distribution distance in each cluster

1. 선택된 클러스터의 수가 정확도에 미치는 영향

Setting Dkl=0.4

If you use only 2,4 clusters like blue one, ~~as the FL server selects the clusters with faster cluster round time~~, the overall FL procedure can be finished in a short time

Moreover, the overfitting problem

Thus, a sufficient number of clusters (e.g., over 20 clusters in this result) should be selected by considering the effect on the accuracy and the round time.

작은 수의 클러스터 수(2,4,5)를 선택할 때, FL서버가 더 빠른 클러스터 라운드 시간을 가지는 클러스터를 선택하기 때문에 전체적인 FL 과정이 짧은 시간에 끝나버린다. 이 경우를 제외하고, 정확도는 **더 많은 클러스터가 선택될수록 정확도는 증가**한다, 20개가 넘는 수가 되면 포화된다.

1. 정확더에 미치는 양자화 수준의 분산

Before 10,000s, no significant discrepancy is observed depending on the distribution of quantization levels. However, after 10,000s, it can be found that the accuracy is gradually increased if a high proportion of quantization levels

0,000초 전에, 양자화 레벨에 따른 분산에는 아무런 차이를 보이지 않는다. 하지만 **10,000초 후, 양자화 레벨이 높은 비율일 수록 정확도가 기하급수적으로 증가**한다.

to reduce the round time while guaranteeing the target accuracy, sufficiently short distribution distances of each cluster should be guaranteed. In addition, a sufficient number of clusters should be selected and suitable distributions of quantization levels should be determined.

요약하자면, 목표하는 정확도를 보증하는 동안 라운드 타임을 줄이기 위해서는, 충분히 짧은 각클러스터 분산 거리를 보증해야한다. 게다가, 충분한 양의 클러스터가 선택되어야하고 적절한 양자화 레벨이 결정되어야한다.

**알고리즘**

Moving on now to SITUA-CQ algorithm with 3 loops and initialization part.

first loop, the algorithm clusters the clients according to their data distributions

Next, in the second loop, the algorithm selects clusters having faster cluster round times until the number of clients and sufficiently high accuracy are satisfied

.In the third loop, the algorithm allocates quantization levels to the selected clusters

These tree step are main point of SITUA-CQ and affect high accuracy.

알고리즘은 3가지 루프와 초기화 부분을 구성하는 SITUA-CQ의 자세한 과정을 보여준다.

첫번째 루프에서, 알고리즘은 클라이언트의 데이터 분산에 따라 클라이언트를 클러스터하고, 각 클러스터의 분포 거리를 IID와 가깝게 유지한다. 다음으로 두번째 루프에서, 알고리즘은 클라이언트의 수를 만족할 때까지 더 빠른 라운드 타임을 가지는 클러스터를 선택한다 그리고 이것은 라운드 타ㅣㅁ이 더이상 연장되지 않을 때 충분히 높은 정확도를 유지하게 한다. 세번째 루프에서, 알고리즘은 양자회 레벨의 미리 예측된 분산을 고려해 선택된 클러스터에게 양자화 레벨을 할당한다. 라운드 시간이 긴 클러스터에 대해 속도를 높이면서도 높은 정확도를 유지합니다.

**결과**

The data here shows that..the accuracy of SITUA-CQ and the comparison schemes

SITUA-CQ can achieve the highest accuracy (i.e., 81.8%) while finishing 500 rounds at 12,425s

First, SITUA-CQ clusters the clients to maintain the distribution distance in each cluster to be close to IID.

Then, SITUA-CQ selects an appropriate number of clusters to avoid overfitting problems

In addition, SITUA-CQ allocates quantization levels to each selected cluster according to the cluster round time

SITUA-CQ와 비교 대상의 정확도를 보여줍니다. SITUA-CQ는 12,425초에 500라운드를 완료하면서 가장 높은 정확도인 81.8%를 달성할 수 있습니다. ITUA-CQ는 클러스터링을 통해 각 클러스터의 분포 거리를 IID에 가깝게 유지합니다. 그런 다음, SITUA-CQ는 과적합 문제와 긴 라운드 시간을 피하기 위해 적절한 수의 클러스터를 선택합니다. 또한, SITUA-CQ는 각 선택된 클러스터에 양자화 수준을 할당합니다. 예를 들어, 특정 클러스터의 라운드 시간이 다른 클러스터보다 상대적으로 길다면, 해당 클러스터에는 낮은 양자화 수준이 할당됩니다.

CLUST가 각 라운드에서 너무 많은 클러스터를 선택하고, FL 서버가 이러한 클러스터에서 모든 클라이언트로부터 업데이트된 딥 모델을 기다리기 때문입니다

SITUA-CQ shows not only comparable performance compared to the optimal solution but also significantly short computation time

Federated learning (FL) is perceived as one of the most promising solutions for wireless networks devices like 6G. So FL is used in various fields such as healthcare, voice recognition, object detection applications, and transport infrastructure.

. SITUA-CQ는 최적 솔루션과 비교하여 비슷한 성능을 보여주며, 상당히 짧은 계산 시간을 가지고 있습니다. 시뮬레이션 결과는 SITUA-CQ가 비교 대상 알고리즘과 비교해 충분히 높은 정확도와 경쟁력 있는 라운드 시간을 보장할 수 있음을 보여줍니다. 게다가, SITUA-CQ는 이질적인 클라이언트가 고려될 때 현저한 개선을 보여줍니다.