紹介論文

Communication-Efficient Learning of Deep Networks from Decentralized Data

H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas

In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017 (AISTATS2017)

Abstract

モバイルデバイスは, モデルを学習するために豊富なデータにアクセスすることができる. それによりデバイス上のユーザビリティが大幅に向上する可能性がある. 例えば, 言語モデルは音声認識とテキスト入力を改善することができ, 画像モデルは自動的に良い写真を選択することができる. しかし, この豊富なデータは多くの場合では, プライバシーに敏感でデータ量が多いため, データセンターへのログ記録やデータセンターへデータを送って学習するといった従来の方法を使用したトレーニングが不可能になることがある. そこで, トレーニングデータをモバイルデバイスに分散したままにし, ローカルで計算された更新を集約して共有モデルを学習するという手法を提案する. この論文では, このような分散アプローチをフェデレーションラーニングと呼ぶことにする. またこの論文は, 反復モデルの平均化に基づくディープネットワークのフェデレーションラーニングのための実用的な方法を提示し, 5 つの異なるモデルアーキテクチャと 4 つのデータセットを考慮して経験的評価を実施する. これらの実験結果は, この設定の定義である不均衡および非 i.i.d なデータ分布に対して堅牢であることを示している.

1 Introduction

スマートフォンやタブレットは多くの人々の主要コンピューティングデバイスである. これらのデバイス (カメラ, マイク, GPS など) の強力なセンサーは膨大な量のデータにアクセスすることができる. そのようなデータで学習されたモデルは, ユーザビリティを大幅に向上させることができるが, データセンターに格納するリスクと責任が大きいことを意味する.

この論文では, データセンターが集中的にデータを格納して保存する必要がなく, 豊富なデータから訓練された共有モデルのメリットを集合的に享受できるようにする学習テクニックを提案する. 各デバイス (クライアント) のデータをデータセンター (サーバ) がまとめるため, フェデレーションラーニングと呼ばれる. 各クライアントには, 決してサーバにアップロードされないローカルトレーニングデータセットがある. その代わりに, 各クライアントは, サーバが保持しているグローバルモデルへの更新を計算し, 更新のみが通信される. ローカルトレーニングデータセットは現在のモデルを改善することにだけ使用されるため, 適用された後は保存する必要がない. このアプローチの主な利点は, モデルトレーニングを生トレーニングデータに直接アクセスする必要性から切り離すことである. また, デバイスとクラウドではなくデバイスのみに攻撃面を限定することでプライバシーとセキュリティのリスクを大幅に削減できる.

Contributions

この論文の主な貢献は以下である.

1. モバイル端末からの分散データをトレーニングする問題を設定
2. フェデレーションラーニングに適用できる簡単で実践的なアルゴリズムの選択

3. 提案手法の経験的評価

具体的には, 各クライアントの SGD(local stochastic gradient descent) とモデル平均化を行うサーバーを組み合わせた FederatedAveraging アルゴリズムを紹介する. この論文は, このアルゴリズムの実験を行い, 不均衡および non-IID なデータに対して堅牢であることを実証し, 分散データのネットワークを訓練するために必要な通信のラウンドを減らすことができることを示す. non-IID で少ない通信での分散学習の手法の提案はこの論文が初めてである.

Federated Learning

フェデレーションラーニングの理想的な問題には, 以下の特性がある.

1. モバイルデバイスから得られる実世界データのトレーニングは, サーバが一般的に利用できるプロキシデータのトレーニングよりもはるかに優れている
2. モバイルデバイスから得られるデータはプライバシーに敏感であり, データサイズが大きいため, データ収集センターに記録しないことが望ましい
3. 教師付きタスクの場合では, データのラベルはユーザーの操作から自然に推論できる.

画像の分類の例では, 将来何回も見られる可能性の高い写真を予測するなどが挙げられる. ただしこのようなデータ (ユーザーがとるすべての写真, パスワード, URL, メッセージ) はプライバシーに敏感である. また, メッセージおよびテキストメッセージにおける言語の使用は, 標準的な言語コーパス (Wikipedia などの文書) と異なる可能性がある. また, 人々が携帯電話で撮った写真は, 一般的な Flickr の写真とはかなり異なる可能性がある. そして, これらの問題のラベルを直接入手可能である. 写真ラベルは彼らの写真アプリを使った自然なユーザーインタラクションによって定義することができる.

Privacy

フェデレーションラーニングはデータセンターでのトレーニングと比較して, プライバシーの点で優れている. 匿名化されたデータセットでさえも, 他のデータとの結合を介してユーザのプライバシーを危険にさらす可能性がある. 対照的に, フェデレーションラーニングのために送信される情報は, モデルを改善するために必要最小限の情報である. また, 生の訓練データよりも多くの情報を含むことは決してなく, 一般にはそれほど多くの情報を含んでいない.

Federated Optimization

フェデレーションラーニングでの最適化問題は Federated Optimization と呼ばれ, 以下のような重要な特性がある.

1. **Non-IID**: クライアントのトレーニングデータは, モバイルデバイスの使用に基づいているため, 特定のユーザーのローカルデータセットは確率分布で表すことができない
2. **Unbalanced**:一部のユーザーは, サービスやアプリを他のサービスよりも多く使用する可能性があるため, さまざまな量のローカルトレーニングデータになる

3. **Massively distributed**:参加するクライアントの数は, 1 クライアントあたりの平均的な例数よりもはるかに多いことが予想される
4. **Limited communication**:モバイルデバイスは, 頻繁にオフラインまたは低速または高価な接続になる

この論文では, Non-IID と Limited communication に重点を置く. クライアント数を K とし, 各ラウンドの始めにクライアントの部分集合 C をランダムに選択する. これらの選択されたクライアントはサーバから現在のグローバルモデルパラメータを受信する. 次に選択されたクライアントはグローバルモデルパラメータとローカルデータセットに基づいてローカルで計算を実行し, サーバに更新結果を送信する. そしてサーバはこれらの更新をグローバルモデルパラメータに適用するという操作が繰り返される.

モデルパラメータを w とし, データを (x_i, y_i) として, 予測損失を $f_i(w) = \ell(x_i, y_i, w)$ とする. クライアント k のデータのインデックス集合を $\mathcal{P}_k$ とする. また全データ数を n として, クライアント k のデータ数を n_k とすると, $|n_k| = \mathcal{P}_k$ となる. この論文は以下の最小化問題を解くことが目標である.

$$f(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w) \text{ where } F_k(w) = \frac{1}{n_k} \sum_{i \in \mathcal{P}_k} f_i(w) \tag{1}$$

2 The FederatedAveraging Algorithm

パラメータ更新に確率的勾配降下 (SGD) が多く用いられている. SGD をフェデレーション環境に適用するには, 各ラウンドで計算を実行するクライアントの割合 C だけ選択し, SGD によって, これらのクライアントが保持するすべてのデータに対する損失の勾配を計算する. つまり, C はグローバルなバッチサイズを制御する値である. $C = 1$, つまりクライアントの選択がフルバッチである場合のアルゴリズムを FederatedSGD（FedSGD）とし, これをベースラインアルゴリズムとする.

学習率 η の FedSGD では, クライアント k は, ローカルデータから SGD をにより, 勾配 $g_k = \nabla F_k(w_t)$ を得る. 次にサーバは以下のようなパラメータ更新を行う.

$$w_{t+1} = w_t - \eta \sum_{k=1}^{K} \frac{n_k}{n} g_k \tag{2}$$

$$\Leftrightarrow \forall k, w_{t+1}^k = w_t - \eta g_k,\ w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k. \tag{3}$$

FedSGD にグローバルなパラメータ更新の前にローカル更新 $w_{t+1}^k = w_t^k - \nabla F_k(w^k)$ を複数回繰り返す操作を加える. これを FederatedAveraging(FedAvg) と定義する. FedAvg のアルゴリズムを Algorithm 1 に示す.(Algorithm 1) E はローカルでのパラメータ更新回数で, B はローカルでのパラメータ更新のバッチサイズである. $W = 1, B = \infty$ の場合, FedSGD となる. n_k 個のローカルデータを持ったクライアント k は 1 ラウンド毎のローカルパラメータ更新回数は $u_k = E\frac{n_k}{B}$ となる.

Figure 1 では, 異なる初期値から訓練された 2 つの MNIST 認識モデルを平均する挙動を観察する. 親モデル w および w_0 は, MNIST のデータセットを用いて, 異なるランダム初期値 (Figure 1, 左) か, 同じランダム初期値 (Figure 1, 右) から学習する. 同じランダム初期化から 2 つのモデルからそれぞれ別々に異なるデータセットを訓練すると, 単純なパラメータの平均化 $\frac{1}{2}w + \frac{1}{2}w_0$ が驚くほどうまく機能することがわかる. また, これらの 2 つのモデルの $\frac{1}{2}w + \frac{1}{2}w_0$ は,2 つのデータセットのいずれかを独立して訓練することによって達成

される最良のモデルよりも完全な MNIST 訓練セットの損失を有意に低下させる. FedAvg の各ラウンドには共通の開始モデル w_t が使用されているので, 同じ直観が得られる.

3 Experimental Results

画像分類と言語モデリングについての FedAvg の有用性を実証する. 学習器として, 隠れ層 2 層の多層パーセプトロン (2NN) と 5×5 の畳み込み層の CNN(CNN) を使用する.

画像分類では MNIST を使用する. 各クライアントの保持するデータを non-IID にするために MNIST では 100 人のクライアントに対し, 正解ラベル毎に並べ替えしたデータを 200 分割し, それぞれのデータを 2 つずつ持つ.

言語モデリングについてはシェークスピアのデータセット [2] を使用し, LSTM によって行内の各文字を読み込んだ後, 次の文字を予測する.. non-IID にするためにシェークスピアから役者毎にクライアントデータセットを作成する. それぞれの役者は少なくても 2 行の文を持っているが, より多くの行を持ってる役者も存在する. これにより 1146 人のクライアントデータセットが作成された.

初めに, クライアント割合を変化させることによってどれほど影響が出るか調査する. Table 1 では,MNIST を 2NN と CNN で学習させて, 目標の accuracy(2NN:97%,,CNN:99%) に達成するまでの通信ラウンド数を記述している.(Table 1)　$B = \infty$ の場合, クライアント数を増やしてもあまり改善されない. $B = 10$ のとき non-IID の場合に有意な改善がみられる.

B を小さくして, E を大きくすることでローカルでの SGD の更新回数が多くなる. Talbe 2 では, 1 回の通信ラウンドで, より多くローカルでの SGD の更新をすると, 通信コストが大幅に減少することがわかる.(Talbe 2)

ただし, E を大きくしすぎると, ローカルデータセットに対して過学習を起こしてしまう可能性がある. 損失関数が convex な場合は初期値に関わらず global minimum に到達するだろう. しかし, non convex な場合は初期値が同じ盆地?(basin) にある限り, 同じ local minimum に収束するかもしれない. Figure 3 では, E を大きくしすぎた場合, 通信繰り返しても, 訓練誤差があまり改善されていないことを示していることがわかる.(Figure 3)

4 Conclusions and Future Work

この論文では, FedAvg がフェデレーション設定で比較的少数のコミュニケーションラウンドでモデルを学習することができることを実験的に示した. フェデレーションラーニングは多くの実用的なプライバシー上の利点がある一方で, 差分プライバシーや安全なマルチパーティ計算を組み合わせることにより強力なプライバシー保証をすることができることは今後の興味深い研究方針である.

References

[1] White House Report. Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy. ournal of Privacy and Confidentiality, 2013.

[2] William Shakespeare. The Complete Works of William Shakespeare. Publically available at

https://www.gutenberg.org/ebooks/100.

[3] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[4] Jakub Konecn y, H. Brendan McMahan, Felix X. Yu, Peter Richtarik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.