Apprendimento distribuito e federato basato su Hypothesis Transfer Learning

Lorenzo Valerio, Andrea Passarella, Marco Conti
IIT-CNR
{l.valerio,a.passarella,m.conti}@iit.cnr.it

Abstract

Attualmente i dati generati da dispositivi all'edge di Internet, quali dispositivi IoT e dispositivi personali degli utenti, vengono analizzati in cloud su data centre remoti. Questo approccio presenta problemi significativi sia dal punto di vista della privacy dei dati, che per il potenziale sovraccarico delle infrastrutture di calcolo e di rete per il trasporto di tali dati Una soluzione particolarmente promettente prevede di spostare la computazione, e di conseguenza l'esecuzione di AI, verso l'edge della rete, dove i dati vengono generati. L'idea infatti è di estrarre conoscenza dai dati tramite metodi di AI utilizzando tecniche di machine learning distribuito e federato. Precisamente, il processo di apprendimento viene eseguito direttamente sui dispositivi che posseggono fisicamente i dati i quali collaborano per apprendere un modello accurato dei dati. In questo articolo, illustriamo una tecnica di apprendimento distribuito basata su Hypothesis Transfer Learning che permette di eseguire un task di apprendimento con performance comparabili a un approccio cloud, riducendo al contempo il traffico di rete generato.

1 Introduzione

Un elemento abilitante per lo sviluppo sostenibile di una Smart City è la capacità di estrarre conoscenza dall'enorme mole di dati prodotti al suo interno sfruttando tecniche di Intelligenza Artificiale (AI). L'AI infatti, unitamente alla presenza dei Big Data, è considerata una tecnologia abilitante che non solo guiderà lo sviluppo tecnologico ed economico dei prossimi anni ma genererà anche un forte impatto dal punto di vista sociale, permeando la vita quotidiana delle persone. Da un punto di vista tecnologico, una tale rivouzione basata sulla AI sarà resa possibile soprattutto grazie alla crescita esponenziale del numero di dispositivi personali e IoT, in grado sia di raccogliere dati dall'ambiente circostante tramite i sensori di cui dispongono, che di crearne di nuovi. Più precisamente, si prevede che entro il 2022 il numero di connesioni a Internet imputabili a dispositivi IoT rappresenteranno più della metà ($\sim 14 * 10^{12}$) di tutte le connessioni che saranno attive globalmente ($\sim 28 * 10^{12}$)[Cisco Systems, 2018a].

Metodologicamente, al momento i dati raccolti dai dispositivi che si trovano all'edge della rete vengono centralizzati su grandi datacenter in cloud dove vengono processati tramite sofisticati algoritmi di AI. Questo approccio, tuttavia, pone diverse questioni connesse alla sua sostenibilità da punto di vista i) della disponibilità delle risorse tecnologiche necessarie, ii) dal punto di vista della privacy dei dati. Si consideri che entro il 2021 il volume di dati generato complessivamente da persone e dispositivi sarà di circa 4 volte quello generato nel 2016 (ovvero 850ZB nel 2021 rispetto a 220 ZB del 2016) e che i cloud datacenter potrebbero avere difficoltà a stare al passo con l'ammontare di dati che verranno generati all'edge della rete da tutti i dispositivi connessi. Nonostante la maggiorparte dei dati prodotti (circa il 90%) non avrà un valenza prolungata nel tempo il restante 10% (circa 85 ZB) avrà un volume che supererà di 4 volte l'attuale traffico gestito e generato dai datacenter (circa 21 ZB all'anno) [Cisco Systems, 2018b]. Inoltre, vanno considerati aspetti di privacy e di accesso ai dati. Precisamente, l'approccio attuale all'AI basato su cloud impone di collezionare e immagazzinare i dati in luoghi fisici che non sono sotto il diretto controllo dei possessori dei dati. Tale aspetto, ha delle ricadute significative in diversi settori come Industria 4.0, Smart Cities, Trasporti ecc.

A fronte di queste considerazioni, un direzione promettente da seguire è di decentralizzare tutto o in parte il processo di AI. Precisamente, l'idea è di spostare la computazione dei modelli di AI verso l'edge della rete, limitando così il flusso di dati che vanno verso il cloud. Precisamente, sfruttando paradigmi come l'edge e il fog computing per eseguire metodi di machine learning distribuito, diventa possibile analizzare ed estrarre conoscenza dai dati senza doverli centralizzare in cloud.

In letteratura, esistono molti metodi di apprendimento distribuito. Molti di questi sono stati sviluppati prendendo come riferimento il contesto dei data-center, dove i)le risorse computazionali e di comunicazione non sono considerate una risorsa limitata e ii) i dati posso essere gestiti opportunamente in modo da soddisfare i tipici requisiti di buon funzionamento degli approcci di Machine Learning (e.g., accesso completo al dataset, dati iid, numero di partizioni del dataset limitato). Tuttavia, tali assunzioni non possono essere ritenute valide per contesti di tipo smart city in cui ci possono essere migliaia di dispositivi eterogenei che raccolgono autonomamente dati dall'ambiente circostante in tempi e locazioni fisiche e

quantità differenti. Inoltre, tipicamente i dispositivi in questi contesti posso disporre di risorse limitate. Servono quindi approcci in grado di operare tenendo presente tali vincoli.

In letteratura sono stati presentati pochi approcci che vanno in questa direzione. Uno di questi, proposto da Google Inc. sotto il nome di Federated Learning [McMahan *et al.*, 2016], propone un meccanismo di apprendimento distribuito in cui ogni dispositivo coinvolto, partecipa all'apprendimento collaborativo di un modello di rete neurale "globale" senza mai spostare i dati dai dispositivi su cui sono salvati. L'approccio di Federated Learning, prevede l'esecuzione distribuita di un variante asincrona dell'algorithmo Stochastic Gradient Descent, in cui si alterna una fase di apprendimento locale portata a termine dai dispositivi coinvolti, seguita da una fase di sincronizzazione globale in cui un parameter server si occupa di aggregare i gradienti ricevuti dai dispositivi per poi comunicare loro questa informazione aggregata. Questo meccanismo si ripete fino a convergenza. In questo articolo, descriviamo una metodologia sviluppata quasi contemporaneamente e in modo indipendente, che condivide gli stessi principi del Federated Learning.

2 Machine Learning Distribuito

In [Valerio *et al.*, 2016] abbiamo proposto un approccio di Machine Learning Distribuito basato su tecniche di Hypothesis Transfer Learning (HTL) che, partendo dalle stesse assunzioni descritte precedentemente, dimostra come sia possibile completare con successo un task di apprendimento distribuito e collaborativo tra dispositivi eterogenei, limitando al contempo il traffico di rete generato. Precisamente, nello scenario considerato, ogni dispositivo detiene un partizione di un dataset su cui addestra un modello (ad es. un classificatore). Successivamente avviene uno scambio di modelli tra i diversi dispositivi coinvolti, i quali eseguono un secondo raffinamento del loro modello applicando un algortimo di HTL. Brevemente, con il Transfer Learning è possibile sfruttare la conoscenza contenuta in uno o più modelli già addestrati (chiamati modelli sorgente) allo scopo di migliorare l'apprendimento di un modello target addestrato su dati (spesso pochi) mai visti dai modelli sorgente. Sebbene l'Hypothesis Transfer Learning sia maggiormente utilizzato per risolvere il problema del cold start (i.e., migliorare l'addestramento di un modello avendo a disposizione pochi dati), in [Valerio *et al.*, 2016] abbiamo dimostrato come esso possa essere utilizzato anche in contesti di apprendimento distribuito.

Nel caso specifico abbiamo usato GreedyTL, un algoritmo di HTL che trova il sottoinsieme ottimale di modelli sorgente che maggiormente contribuiscono all'addestramento del modello locale (modello target). Si noti che nel nostro approccio, ogni dispositivo è sia contributore che fruitore di modelli sorgente nei confronti degli altri dispositivi. L'algoritmo proposto si compone di una prima fase in cui i device apprendono un modello locale, che viene successivamente condiviso con tutti. Nella seconda fase viene applicato GreedyTL allo scopo di raffinare i modelli locali. Infine i modelli raffinati vengono aggregati in un unico modello globale. In questo modo è possibile i) migliorare il modello appreso localmente integrando informzioni contenute in altri modelli e ii) ottenere, alla fine del processo, un modello globale più accurato dei singoli modelli provenienti dai singoli dispositivi.

3 Risultati e conclusioni

Questo approccio, è stato testato su diversi task di apprendimento come l'activity recognition (HAPT Dataset), il riconoscimento di numeri scritti a mano (MNIST Dataset), la classificazione di quali classi di alberi ricoprono determinate aree di foresta (Covertype Dataset). In ognuno di questi task si è dimostrato efficace dal punto di vista dell'accuratezza nelle predizioni, ma sopratutto efficiente nella comunicazione. Infatti, prendendo il caso dell'activity recognition, abbiamo dimostrato il nostro approccio riusciamo a risparmiare più del 70% del traffico di rete, rispetto ad una soluzione centralizzata in cui tutti i dati vengono raccolti in cloud, ottenendo un'accuratezza del 97% (solo 2% inferiore al cloud). In aggiunta, tramite ulteriori sperimentazioni abbiamo potuto appurare anche la robustezza del nostro metodo in presenza di modelli sorgente corrotti da rumore casuale. Sfruttando le caratteristiche di GreedyTL siamo in grado di selezionare i modelli più informativi, scartando, di conseguenza quei modelli che non apportano nessuna informazione al processo di learning [Valerio *et al.*, 2017].

Le competenze presentate sono state sviluppate all'interno delle seguenti attività progettuali H2020: REPLICATE (691735) , SoBigData (654024), and AUTOWARE (723909).

Riferimenti bibliografici

[Cisco Systems, 2018a] Cisco Systems. White paper: Cisco visual networking index: Forecast and trends, 2017–2022. pages 2017–2022, 2018.

[Cisco Systems, 2018b] Inc. Cisco Systems. Cisco Global Cloud Index: Forecast and Methodology, 2016–2021. *White Pap.*, page 46, 2018.

[McMahan *et al.*, 2016] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. 54, 2016.

[Valerio *et al.*, 2016] Lorenzo Valerio, Andrea Passarella, and Marco Conti. Hypothesis Transfer Learning for Efficient Data Computing in Smart Cities Environments. In *2016 IEEE Int. Conf. Smart Comput. SMARTCOMP 2016*, 2016.

[Valerio *et al.*, 2017] L. Valerio, A. Passarella, and M. Conti. A communication efficient distributed learning framework for smart environments. *Pervasive Mob. Comput.*, 41, 2017.