Martin Georgiu, *Kolaborativní učení (Federated learning)*

(Posudek oponenta na bakalářskou práci)

Práce se zabývá přístupem k strojovému učení (machine learning, ML) nazývaným kolaborativní učení (Federated learning, FL). Zaměřuje se na metody trénování ML modelů v distribuovaném prostředí s důrazem na soukromí. Data zůstávají u klientů a autor modelu k nim nemá přímý přístup.

Po stručné úvodní kapitole autor popisuje základní pojmy a framework kolaborativního učení v kapitole 2, na což navazuje popis existujících instancí kolaborativního učení v kapitole 3. Kapitola 4 se stručně věnuje problému bezpečnosti a v kapitole 5 pak autor popisuje experimenty s učením neuronové sítě pro rozpoznávání kožních problémů (nejen) pomocí kolaborativního učení.

Z textu práce je patrné, že autor nastudoval nejen existující literaturu k tomuto tématu, ale také existující knihovny, které popisované algoritmy implementují. Autor také naprogramoval a zveřejnil ukázkový příklad, jak pomocí těchto knihoven natrénovat neuronovou síť pro rozpoznávaní kožních problémů ze snímků kůže. K samotnému textu práce mám ale značné výhrady, a to jak k jeho struktuře, tak k jeho stylistice. Provedení a vyhodnocení experimentů také není bez problémů.

Práce je psaná anglicky, což oceňuji. Nicméně text je na více místech příliš neformální a vágní. Také obsahuje větší množství gramatických chyb a stylisticky pochybných konstrukcí, které brání plynulému čtení textu (str. 12, "But if we would stick to the premise of cross-device."). Nejvíce mi v práci chybí širší úvod do problematiky strojového učení a učení neuronových sítí. Mnoho pojmů a konceptů, které se v práci objevují, nejsou vysvětleny vůbec nebo jen velmi vágně (hyperparameters, epochs, batch norm layers, loss function, loss value, optimizer, TensorBoard, Focal loss + γ value, ...). Přidání obecného úvodu do učení neuronových síti by čtenáři velice usnadnilo pochopení výhod a nevýhod kolaborativního učení jako celku i jednotlivých variant představených v této práci.

Také představení FL v kapitole 2 je dle mého názoru nedostačující a její struktura matoucí. Některé části této kapitoly (Sekce 2.1, 2.2) patří spíše do úvodní kapitoly, zatímco sekce 2.4 by měla následovat až po představení obecného FL frameworku (sekce 2.5). Hlavní myšlenka FL, zachycena v obrázcích 2.1 a 2.2, by si zasloužila větší prostor než aktuální skromný text v sekci 2.5. Co se týče jiných přístupů (sekce 2.6), není jasné, zda se jedná o alternativy k FL nebo o rozšíření či úpravu FL frameworku.

Podobně kapitola 4 působí nesourodým dojmem. Zatímco v úvodu je kladen důraz na dva typy útočníků, kteří se snaží získat informace o soukromých datech, ve zbytku kapitoly (sekce 4.1–4.3) již tyto dva typy útočníků nehrají žádnou roli. Sekce 4.2 se dokonce zabývá problémem ochrany natrénovaného modelu a ne ochrany dat. Podle popisu v textu jsou tyto problémy (inference attacks, stealing model) a techniky pro ochranu dat (differential privacy) obecné, chybí zde zasazení do kontextu FL nebo problémy a řešení specifické pro FL.

Co se týče implementace a vlastních experimentů, zde oceňuji detailní rozbor problému (nevyvážená sada dat) i způsobů řešení. Naopak chybí detailní popis datové sady i vybrané architektury neuronové sítě. Autor porovnává 4 konfigurace kolaborativního učení s klasickým učením vybrané neuronové sítě. Zde je podivuhodné, že každé konfiguraci bylo přiděleno různé celkové množství epoch. Očekával bych, že každá konfigurace bude trénovaná stejný počet epoch. Vysvětlení toho, jak byl určen počet epoch pro různé konfigurace v textu chybí. Na obrázku 5.5 konfigurace číslo 3 končí po 25 epochách, ale podle obrázku 5.4 mělo trénování trvat $25 \times 5 = 125$ epoch. Závěry z porovnání konfigurací 2 a 3, a konfigurací 2 a 4 dávají smysl vzhledem k obrázku 5.5, ale ne pro výsledné hodnoty přesnosti uvedené na obrázku 5.4. V sekci 5.2.3 autor nastiňuje zajímavou možnost klást při trénovaní větší důraz na eliminaci falešně negativních klasifikací. Je škoda, že není této myšlence věnován větší prostor a že tato možnost nebyla v experimentech prozkoumána, jelikož by dle mého názoru šlo o zajímavý výsledek.

Další otázky, které vyvstaly při čtení práce, jsou:

- str. 6, Longevity: Proč je dle autora problematické v decentralizovaném prostředí aktualizovat existující model? Jaké nové problémy tomu brání v porovnání s trénováním nového modelu?
- Sekce 3.2.3: Co reprezentuje parameter μ? Je stejný pro všechny klienty?
- Sekce 3.3: Jsou "batch norm" vrstvy fixní v architektuře sítě, nebo si je každý klient doplní samostatně? Je-li pravda první možnost, jak pak vypadá finální model, když klienti neposílají aktualizace pro tyto vrstvy na server?
- Sekce 4.1.2: Jaké vlastnosti datových sad se útočník snaží odhalit?
- Sekce 4.2: Jaké jsou možnosti obrany proti útočníkovi, který sa snaží ukrást samotný model?

- Proč není implemantace součastí odevzdání jako příloha práce?

Dle mého názoru by tato práce měla potenciál být výbornou, pokud by autor přidal obšírnější úvod do problematiky, lépe vysvětlil používané pojmy a více propojil představené problémy a techniky řešení (s vysvětlením jejich výhod a nevýhod). Ale v současné podobě, vzhledem k výše popsaným problémům (a dalším typografickým nepřesnostem), hodnotím práci pouze známkou **dobře**.

V Praze dne 7. 6. 2023

Mgr. Martin Blicha, Ph.D.
oponent práce