dr hab. inż Jan Chorowski
Pathway

Wrocław, 03.03.2024

Recenzja rozprawy doktorskiej mgr Jakuba Świątkowskiego pt. „Variational Inference Applications in Deep Learning"

1 Tematyka rozprawy

Dorobek Doktoranta przedstawiony w rozprawie doktorskiej składa się z czterech powiązanych tematycznie artykułów dotyczących zastosowań metod wariacyjnych w uczeniu głębokim. Przedstawione prace dopełniają się i wykazują biegłość doktoranta w teoretycznych i praktycznych aspektach stosowania głębokich sieci neuronowych. Pierwsze dwie prace dotyczą Bayesowskiego uczenia sieci neuronowych w którym celem jest wyznaczenie rozkładu prawdopodobieństwa parametrów sieci. Kolejne dwie prace używają metod wariacyjnych w uczeniu auto-enkoderów zastosowanych do modelowania prozodii w systemach syntezujących mowę.

2 Ocena rozprawy

Rozprawę czytałem z duża przyjemnością. Na szczególną uwagę zasługują pierwsze rozdziały pracy które zwięźle i precyzyjnie definiują niezbędne pojęcia. Bardzo ucieszyła mnie dbałość o przedstawienie zarówno aspektów teoretycznych dot. uczenia sieci, jak również syntezy mowy. Nawet w pracach empirycznych, nastawionych na poprawę jakości działania sieci widać chęć Autora do świadomego działania podpartego aspektami teoretycznymi, a nie jedynie przeszukiwanie przestrzeni hiper-parametrów. Tę dążność Doktoranta do zrozumienia teoretycznego uczenia sieci neuronowych dobrze ilustruje praca "How Good is the Bayes Posterior in Deep Neural Networks Really?" stanowiąca podstawę rozdziału 5, której celem jest wytłumaczenie rozbieżności w teorii i praktyce stosowania sieci Bayesowskich.

O jakości pracy świadczy również ranga konferencji na których zostały przedstawione uzyskanie przez doktoranta wyniki: ICML będąca w ścisłej czołówce konferencji dotyczących uczenia maszynowego oraz Interspeech uznawana za wiodącą konferencję dotyczącą przetwarzania mowy.

2.1 Ocena treści

Rozdziały wprowadzające 1-3.

Pierwsze 3 rozdziały pracy wprowadzają czytelnia w tematykę rozprawy. W rozdziale pierwszym zwięźle opisano czym są głębokie sieci neuronowe, oraz przedstawiono niezbędne aspekty metod wariacyjnych: wyprowadzono Evidence Lower Bound (ELBO) oraz opisano możliwość optymalizacji ELBO przez reparameterization trick.

W rozdziale 2 wprowadzono sieci Bayesowskie oraz opisano dwie uczenia: obliczając rozkład prawdopodobieństwa a posteriori z użyciem opisanych w rozdziale 1 metod wariacyjnych oraz próbkowanie rozkładu a posteriori metodą MCMC z wykorzystaniem dynamik Langevina. Metody te zostały wykorzystane w rozdziałach 4 i 5 pracy.

W rozdziale 3 przedstawiono podstawy przetwarzania mowy oraz opisano obecnie popularne architektury głębokich siec neuronowych wykorzystywanych w syntezie mowy. Wprowadzono również autoenkodery wariacyjne wykorzystane do poprawy syntezy w rozdziałach 6 i 7.

Rozdziały wstępne są przystępnie napisane i wskazują na biegłość Doktoranta w aspektach teoretycznych ucznia głębokich sieci neuronowych.

Rozdział 4 "k-tied Normal Distribuion".

Rozdział przedstawia pierwszą z prac składających się na dorobek doktoranta. W pracy zaobserwowano że możliwa jest redukcja ilości parametrów rozkładów normalnych przybliżających rozkład a posteriori dla wyuczonej sieci. Typowo, w metodach wariacyjnych każdy parametr sieci opisany jest rozkładem normalnym z właściwymi dla danego parametru średnią i wariancją. W pracy zauważono, że wariancje wag sieci należących do jednej warstwy mają swoistą strukturę i możliwe jest przybliżenie ich macierzą niskiego rzędu. Wymuszenie takiej kompresji wariancji zmniejsza ilość parametrów niezbędnych do wyznaczenia podczas uczenia sieci, przyspiesza zbieżność oraz poprawia dokładność uczonych sieci.

Przedstawione wyniki mają znaczenie praktyczne. Osobiście często stosowałem regularyzacje sieci neuronowych polegających na zaszumianiu wag, w tym również dobierając szum metodami wariacyjnymi. Jest to bardzo skuteczna metoda regularyzacji i jej poprawa jest mile widziana.

Opisane w rozdziale 4 przybliżanie wariancji nisko-wymiarowymi macierzami bardzo przypomina metody ograniczania parametrów utrzymywanych przez optymalizatory, jak np. Adafactor[1] który faktoryzuje macierze drugich momentów, ograniczając ilość statystyk utrzymywanych przez optymalizator. Ciekawią mnie podobieństwa między tymi metodami i określenie czy można wykorzystać nisko-wymiarową strukturę macierzy do podzielenia na grupy wag w każdej warstwie.

Rozdział 5 "How Good is the Bayes Posterior".

W rozdziale przedstawiono próbę pogodzenia teorii uczenia sieci neuronowych z wynikami praktycznymi. Typowo, podczas stosowania Bayesowskich sieci neuronowych wprowadza się poprawkę zmniejszającą wariancję rozkładu a posteriori (wprowadza się mnożnik zwany temperaturą). Działanie to nie ma uzasadnienia teoretycznego i poparte jest jedynie aspektami empirycznymi. W

[1] https://arxiv.org/abs/1804.04235

rozdziale przedstawiono szereg hipotez mających wyjaśnić tę rozbieżność. Choć nie osiągnięto jednoznacznych konkluzji, przedstawiono hipotezy wskazujące na ograniczenia obranych rozkładów a priori.

Rozdział 5 był dla mnie najciekawszym fragmentem rozprawy. Opisana nadmierna spójność losowo inicjalizowanej sieci skłania do ponownego przyjrzenia się regułom losowania wag sieci i ich wpływowi na uczenie się modeli.

Doktorant zaznacza, że nie jest głównym autorem pracy przedstawionej w rozdziale 5, był jednak głównym pomysłodawcę zbadania konieczności wyostrzania rozkładu a posteriori. Umiejętność wskazywania ciekawych tematów badań jest cenną umiejętnością posiadaną przez Doktoranta.

Rozdziały 6 i 7 Modelowanie prozodii w syntezie mowy na potrzeby dubbingu.

W rozdziałach 6 i 7 przedstawiono wyniki dwóch prac opublikowanych na konferencji Interspeech. Tematem prac jest system syntezy mowy na potrzeby dubbingu. Celem jest synteza przetłumaczonej mowy przy zachowaniu cech głosu oraz intonacji oryginalnego nagrania. Wprowadzono szereg zmian do bazowego modelu VITS, w tym:

- globalne (rozdział 6) i lokalne (rozdział 7) enkodery prozodii,
- odszumianie,
- uwarunkowanie modelu na cechach mówców,
- uwarunkowanie modelu na cechach języka docelowego.

Do zakodowania prozodii wykorzystano auto-enkodery wariacyjne, w których rozkład zmiennych ukrytych optymalizowany jest przez minimalizację kryterium ELBO wprowadzonym w rozdziale 1. Wykazano, że modele wzbogacone o informacje prozodyczne skutkują lepszą jakością dubbingu w testach odsłuchowych.

Rozdziały 6 i 7 stanowią domknięcie praktyczne rozprawy, dopełniając bardziej teoretyczne rozważania przedstawione w rozdziałach 4 i 5. Na uwagę zasługuje stopień skomplikowania zaproponowanych systemów syntezy mowy, wskazujący na biegłość doktoranta w stosowaniu głębokich sieci neuronowych.

3 Konkluzja końcowa

Osiągnięcia Autora to:

- wprowadzenie k-związanych rozkładów normalnych usprawniających uczenie Bayesowskich sieci neuronowych,
- próba wytłumaczenia konieczności wyostrzania rozkładów prawdopodobieństwa a posteriori dla Bayesowskich sieci neuronowych,
- usprawnienie sieci syntezujących mowę o wariacyjne uczenie się reprezentacji prozodii za pomocą auto-enkoderów wariacyjnych.

Za najciekawszą część pracy uważam rozdział 5 i próbę wyjaśnienia rozbieżności między teorią a praktyką uczenia sieci neuronowych. W praktyce stosowania sieci neuronowych często względy praktyczne przeważają i stosowane jest wiele trików nie mających poprawnego uzasadnienia. Próby wyjaśnienia potrzeb stosowania tych trików porządkują wiedzę i pozwalają na lepsze zrozumienie istotny procesów uczenia się sieci neuronowych.

Dorobek Autora wykazuje biegłość w zagadnieniach teoretycznych i praktycznych uczenia głębokich sieci neuronowych metodami wariacyjnymi. Ponadto Autor umie stawiać hipotezy oraz eksperymentalnie je weryfikować.

Podsumowując, stwierdzam, że przedłożona mi do recenzji rozprawa p. mgra Jakuba Świątkowskiego spełnia z nawiązką wymagania stawiane w Ustawie rozprawom doktorskim i wnoszę o jej dopuszczenie do publicznej obrony.

Jan Chorowski