POLSKO-JAPOŃSKA WYŻSZA SZKOŁA TECHNIK KOMPUTEROWYCH

ポーランド日本情報工科大学

Warszawa, 19 stycznia 2011 r.

prof. dr habil. Witold Kosiński
Polsko-Japońska Wyższa Szkoła
Technik Komputerowych, Warszawa
Uniwersytet Kazimierza Wielkiego
Bydgoszcz

Opinia na temat rozprawy doktorskiej mgr. inż. Łukasza Brockiego:

Koneksjonistyczny model języka w systemach rozpoznawania mowy

Niniejszą recenzję przygotowałem na zlecenie Rady Wydziału Informatyki Polsko-Japońskiej Wyższej Szkoły Technik Komputerowych, która prowadzi przewód doktorski mgr. inż. Łukasza Brockiego. Promotorem rozprawy jest dr hab. inż. Krzysztof Marasek, prof. PJWSTK.

Uwagi wstępne

Jednym z podstawowych wyzwań stawianych współczesnie w technice jest ułatwienie komunikacji między ludźmi mówiącymi różnymi językami oraz między człowiekiem a maszyną, w szczególności komputerem. To wyzwanie stawiane jest specjalistom od języka, lingwistom, ale także przed informatykami zajmującymi się przetwarzaniem języka naturalnego i jego automatycznym rozpoznaniem oraz syntezą.

Działem nauki, który obejmuje te zagadnienia swoim zainteresowaniem jest sztuczna inteligencja oraz jej informatyczna część – inteligencja obliczeniowa. Jednym z najlepiej rozwiniętych narzędzi inteligencji obliczeniowej są sztuczne sieci neuronowe. Klasyfikacja sieci podlega różnym kryteriom: budowa, przepływ sygnału oraz metody ich adaptacji – uczenia. Kryterium przepływu sygnału dzieli sieci na jednokierunkowe oraz sieci rekurencyjne. W sieciach jednokierunkowych przepływ sygnałów następuje w jednym kierunku: od wejścia do wyjścia. Najczęściej sieci jednokierunkowe mają budowę warstwową, gdzie warstwy są uporządkowane: mamy warstwę wejściową, układ warstw ukrytych i warstwę wyjściową. Wtedy sygnał przepływa od warstwy wejściowej poprzez warstwy ukryte do warstwy wyjściowej.

Sieci rekurencyjne różnią się od sieci jednokierunkowych, gdzie przetwarzany sygnał jest po wstępnym przetwarzaniu podawany jeszcze raz na wejście do sieci. W szczególności w warstwowych sieciach rekurencyjnych jest to realizowane sprzężeniem zwrotnym z warstwą ukrytą. Cechą wyróżniającą

ul. Koszykowa 86, 02-008 Warszawa; tel.: +48 22 58 44 500, fax: +48 22 58 44 501
e-mail: pjwstk@pjwstk.edu.pl, www.pjwstk.edu.pl
Konto PJWSTK: 38 2190 0002 3000 0046 2893 0101, Bank DnB NORD Polska S.A.

sieci rekurencyjne jest ich stosunkowo wyższa zdolność adaptacji, w porównaniu z sieciami jednokierunkowymi. Niestety ceną za to jest dłuższy proces uczenia.

Sprzężenie zwrotne realizowane przez połączenia rekurencyjne tworzą w tych sieciach pamięć krótkotrwała. I to właśnie okazało się dla autora recenzowanej pracy doktorskiej drogocenne. Zbudowany przez Niego koneksjonistyczny model języka polskiego można było z powodzeniem zrealizować właśnie w takim typie sieci i zbudować działający w systemie rzeczywistym system rozpoznawania mowy z dużym słownikiem.

Zawartość rozprawy

Rozprawa składa się z 9 rozdziałów, literatury, która zawiera 55 pozycji oraz dodatku z przykładowymi zdaniami testowymi. Praca liczy 95 stron. Praca zwiera streszczenia w języku polskim i angielskim

Rozprawa rozpoczyna się Wprowadzeniem zawierającym sformułowanie problemu, tezy i układ pracy.

Układ pracy podzielonej na rozdziały został tak opracowany, aby zbalansować informacje o zagadnieniach związanych z modelowaniem języka na potrzeby rozpoznawania mowy i zadbać o w miarę przejrzysty opis aktualnego stan badań i na jego tle przedstawić własny wkład Autora do prac w danej dziedzinie. Wybrany przez Doktoranta układ pracy jest, według mnie, bardzo dobry, ponieważ pozwala zorientować się w danej tematyce, a dopiero na samym końcu każdego rozdziału Autor przedstawia swoją koncepcję rozwiązania opisanych wcześniej problemów. Zgadzam się z Doktorantem, że taki układ pracy zapewnia tematyczną spójność i przejrzystość pracy.

W rozdziale 2 zaprezentowano ogólny schemat działania automatycznych systemów rozpoznawania mowy. Autor rozpoczyna swój wywód od definicji procesu rozpoznawania mowy jako znalezienia optymalnej sekwencji słów, która będzie miała największe prawdopodobieństwo względem modelu akustycznego i modelu języka. W kolejnych częściach rozdziału Doktorant przedstawia metody przetwarzania sygnałów cyfrowych, sposoby parametryzacji sygnału, podstawowe pojęcia fonetyczne, a także tradycyjne n-gramowe modele języka.

Rozdział 3 jest najobszerniejszy i podzielono go na 13 podrozdziałów. W tym rozdziale, po uprzednim wprowadzeniu do tematyki sieci neuronowych, przedstawiono zasadę działania koneksjonistycznych modeli języka. Autor opisał powszechnie występujący problem znany pod nazwą klątwy wielowymiarowości, która polega na tym, że model języka musi przypisywać prawdopodobieństwa dla wszystkich możliwych kombinacji słów, a ilość możliwości

zwiększa się wykładniczo wraz z ilością modelowanych słów i długością historii modelu języka. W tym rozdziale opisano również sposoby generalizacji wiedzy przez tradycyjne n-gramowe i koneksjonistyczne modele języka. Na podkreślenie zasługuje fakt, że w odróżnieniu od tradycyjnych modeli języka, sieci neuronowe w sposób automatyczny znajdują zależności i podobieństwa między słowami. Zapewnia to rozproszona reprezentacja słów. W kolejnych podrozdziałach znajduje się formalna definicja koneksjonistycznego modelu języka, jak i sposoby trenowania takich modeli. Z uwagi na fakt, że sieci neuronowe modelujące język są duże, to wymagają wyjątkowo dużo mocy obliczeniowej. Autor opisuje eksperymenty innych badaczy, którzy trenowali sieć neuronową składającą się z aż 17 964 neuronów w warstwie wyjściowej. Wykorzystywali oni do trenowania klaster 32 jednakowych serwerów. Doktorant nie miał dostępu do tak takiej ilości jednakowych komputerów i zaproponował swój własny sposób trenowania modeli języka, który może korzystać z klastrów komputerów o różnych prędkościach. Rozdział 3 kończy się porównaniem n-gramowych i koneksjonistycznych modeli języka na przykładzie języka angielskiego.

Rozdział 4 poświęcony jest koneksjonistycznym modelom języka z nieograniczonym kontekstem, które realizowane są przez rekurencyjne sieci neuronowe. W przeciwieństwie do jednokierunkowych sieci neuronowych, rekurencyjne sieci pozwalają na uchwycenie dynamiki zjawisk nieosiągalnej dla tych pierwszych. Dodatkowo Doktrant w swoich eksperymentach wykorzystuje sieci znane pod nazwą *Long Short-Term Memory* (LSTM), które posiadają specyficzną topologię, zapewniającą jeszcze wyższą skuteczność działania od tradycyjnych, rekurencyjnych sieci neuronowych. Doktorant cytuje szereg prac, które udowadniają wyższość tej nowej topologii nad innymi. Głównym osiągnięciem tego rozdziału 4 jest porównanie, na przykładzie języka polskiego, n-gramowego modelu języka, jednokierunkowej i rekurencyjnej sieci neuronowej.

Do celów eksperymentu wybrano korpus ze stenogramów z posiedzeń kadencji Sejmu czwartej kadencji. Sieć neuronowa jednokierunkowa składała się z 1 066 408 wag synaptycznych. W warstwie wyjściowej zastosowano funkcję aktywacji softmax. Model był trenowany algorytmem gradientowym, z dodatkowym składnikiem momentowym. Współczynnik uczenia wynosił: 10^{-7}, a współczynnik momentowy ustawiono na 0.99. Parametry ustawiono empirycznie. W pierwszym eksperymencie wykorzystano rekurencyjną sieć neuronową trenowaną algorytmem *Backpropagation Through Time* (BPTT). Pozostałe parametry trenowania pozostały takie same, jak przy użyciu jednokierunkowej sieci neuronowej. Trenowanie modelu wykorzystującego sieć jednokierunkową zajęło 16 dni, a trenowanie sieci rekurencyjnej 12 dni. Eksperymenty wykonane przez Autora potwierdzają, że rekurencyjna sieć neu-

ronowa uzyskała najlepszy wynik, co potwierdza pierwszą tezę pracy.

Rozdział 5 opisuje sposób implementacji modelu języka w systemach rozpoznawania mowy. Kluczowym czynnikiem jest opisana przez Doktoranta faktoryzacja prawdopodobieństw obliczanych przez model języka na drzewie leksykalnym, które jest strukturą danych wymaganą do rozpoznawania mowy. Według Autora tradycyjny, algorytmiczny sposób faktoryzacji jest zbyt kosztowny obliczeniowo, aby go zaimplementować w systemach rozpoznawania mowy działających w czasie rzeczywistym. Doktorant przedstawia własną metodę trenowania sieci neuronowej w taki sposób, aby nie obliczała pełnego prawdopodobieństw występowania słów lecz ich zfaktoryzowane prawdopodobieństwo, umieszczone w węzłach sieci leksyklanej. Autorski sposób faktoryzacji prawdopodobieństwa jest głównym osiągnieciem Doktoranta. Pozwala on na znaczną redukcję potrzebnych obliczeń. Rozdział kończy się eksperymentem porównującym jakość i prędkość systemu rozpoznawania mowy wykorzystującego algorytmiczną faktoryzację i sieć neuronową obliczającą bezpośrednio zfaktoryzowane prawdopodobieństwo. Drugie podejście realizuje funkcję rozpoznawania mowy w czasie rzeczywistym i jest - w ocenie Doktoranta - 33 razy szybsze od pierwszej wersji. Wyniki niniejszego eksperymentu potwierdzają drugą tezę rozprawy doktorskiej.

Ocena wyników rozprawy

Największym problemem w automatycznym rozpoznawaniu mowy jest długotrwały proces przeszukiwania hipotez ze względu na wielkość przestrzeni reprezentującej słowa i ich sekwencje. Dla przyspieszenia tego procesu wymagane są odpowiednie struktury danych, oraz algorytmy przeszukiwania i ich możliwa optymalizacja. Stąd w pierwszych systemach rozpoznawania zastosowano liniowy leksykon, który reprezentował liniową sekwencję fonemów niezależną od transkrypcji fonetycznej innych słów. Ich łatwość implementacji niestety łączyła się z większa przestrzenią przeszukiwania, ponieważ liniowe leksykony nie uwzględniały podobieństw fonetycznych na początkach słów. Znacznie bardziej efektywnym leksykonem jest taki, który wykorzystuje strukturę danych zwaną drzewem leksykalnymi, gdzie dokonuje się kompresji części fonemów, które są wspólne dla różnych słów, tym samym istnieje możliwość zmniejszenia przestrzeni przeszukiwania. W drzewie tym korzeń jest miejscem, w którym zaczyna się proces wyszukiwania hipotez. Kolejne węzły reprezentują fonemy. Liście drzewa zawierają wskaźniki do odpowiednich słów w słowniku. Rozpoznawanie przebiega poprzez testowanie hipotez uaktywniając niektóre węzły drzewa. Wtedy w węzłach aktywnych znajdują się hipotezy słów. Większość hipotez znajduje się zawsze na początku

drzewa leksykalnego, gdyż to wynika z pojawieniem się granic słów podczas procesu wyszukiwania. W ten sposób drzewo leksykalne kompresuje wspólne początki transkrypcji fonetycznej słów.

Kompresja wspólnych początków (prefiksu) jest wielką zaletą drzewa, ale jednocześnie jego wadą: gdyż pojedyńczy węzeł, przez to, że jest częścią więcej niż jednego prefiksu, jest automatycznie częścią więcej niż jednego słowa. Dopiero węzły końcowe, tzn. liście drzewa, reprezentują pojedyncze słowa. A to oznacza, że będąc w węźle aktywnym nie można stwierdzić do jakiego słowa należy hipoteza, która znajduje się w tym (jak i i każdym innym) węźle drzewa leksykalnego, poza jego liściem. Dopiero, gdy hipoteza dotrze do końca drzewa można jednoznacznie stwierdzić, jakie słowo reprezentuje. W przeciwieństwie do drzewa leksykalnego, leksykon linowy jasno definiuje, który węzeł należy do którego słowa. Pozwala to na nałożenie prawdopodobieństwa modelu języka na samym początku słowa, a nie na jego końcu, jak ma to miejsce w drzewie leksykalnym. Wcześniejsze nałożenie prawdopodobieństwa modelu języka pozwala na przyśpieszenie i polepszenie jakości procesu rozpoznawania mowy, ponieważ na wcześniejszym etapie można odrzucić mniej obiecujące hipotezy.

I tutaj pojawia sie oryginalna metoda faktoryzacji prawdopodobieństw zaproponowana przez Doktoranta. Polega ona na rozłożeniu pojedynczej wartości na czynniki, które znajdują się w rozgałęzieniach drzewa leksykalnego. Algorytm wybierający najbardziej prawdopodobne hipotezy wymnoży poszczególne czynniki, tak aby docierając do liścia finalne prawdopodobieństwo sekwencji stanów było równe pierwotnej wartości, która znajdowała się w liściu drzewa. To stanowi podstawę budowanego koneksjonistycznego modelu języka, który składa się ze słownika zawierającego cechy słów, oraz sieci neuronowej wykorzystującej w warstwie wyjściowej tyle grup definiowanych przez funkcje aktywacji softmax, ile poziomów ma drzewo leksykalne. To pozwala na dekompozycje problemu. W jego ramach oblicza się bezpośrednio zfaktoryzowane względem drzewa leksykalnego prawdopodobieństwa występowania słów. Istotne jest, że węzły drzewa leksykalnego, które znajdują się w rozgałęzieniach drzewa, symbolizują równocześnie neurony warstwy wyjściowej. Każdy z neuronów odpowiada za wyliczane sfaktoryzowanego prawdopodobieństwa modelu języka.

Rozprawa dowodzi dużej pomysłowości i biegłego opanowania przez Doktoranta współczesnych modeli statystycznych języka, teorii sieci rekurencyjnych i metod ich adaptacji. Zaproponowane przez Doktoranta metoda faktoryzacji prawdopodobieństw pozwala zwiększyć szykość automatycznego rozpoznawania języka. Paradoksalnie wynik ten osiągnięto posługując sie w zasadzie 1-gramowym (unigramowym) modelem języka. Ten element należy do oryginalnych osiągnięć Doktoranta. Należy podkreślić umiejetności imple-

mentacyjne Doktoranta i sprawne posługiwanie się złożonymi algorytmami adaptacji rekurencyjnych sieci neuronowych. Należy też podkreślić twórcze zastosowanie jednego z głównych narzędzi inteligencji obliczeniowej.

Na zakończenie tej części recenzji stwierdzam, że rozprawa mgr. inż. Łukasza Brockiego zawiera oryginalny dorobek naukowy Doktoranta, a jej wyniki są wartościowe. Ponadto wnoszą do zastosowań sieci neuronowych w zagadnieniach technicznych nowe horyzonty badań. Należy na zakończenie tego punktu stwierdzić, że załączone wyniki eksperymentów wskazują, że tezy zostały wykazane.

Uwagi krytyczne i dyskusyjne

1. Wyraźnie odczuwa się brak w prezentacji konkretnych zależności, funkcji równań, które były podstawą budowanych algorytmów i ich implementacji. Tak, że recenzent musi wykazać się dobrą wolą i dużym poziomem zaufania do uzyskanych wyników. Co prawda miałem okazję być na prezentacji Doktoranta zbudowanego systemu, ale dalej odczuwam pewien niedosyt.

2. Doktorant (np. str. 47, ostatni akapit) dla obiektów policzalnych używa terminu ilość możliwych sekwencji, zamiast liczba możliwych.

3. Jak konstruuje Doktorant algorytm adaptacji wag synaptycznych w procesie sprzężenia zwrotnego i jaka jest dokładnie postać funkcji aktywacji zaznaczonych schematycznie na Rys.5 na str.53?

4. Interpunkcja (w szczególności użycie przecinków) stosowana przez Doktoranta wymaga poważnej korekty.

Uwagi końcowe

Przesłana do mnie do recenzji rozprawa doktorska Łukasza Brockiego p.t.: Koneksjonistyczny model języka w systemach rozpoznawania mowy promotorstwa prof. dr habil. Krzysztofa Maraska, spełnia wymogi stawiane rozprawom doktorskim (Ustawa o stopniach naukowych i o tytule naukowym oraz o stopniach i tytule w zakresie sztuki z dnia 14 marca 2003 roku, Dziennik Ustaw Nr 65, poz. 595 wraz z późniejszymi zmianami) w dziedzinie nauk technicznych w dyscyplinie informatyka. W związku z tym wnioskuję o dopuszczenie Doktoranta mgr. inż. Łukasza BROCKIEGO do publicznej obrony.

Witold Kosiński