Częstochowa, dn. 23 października 2018 r.

Dr hab. inż. Rafał Scherer, prof. Politechniki Częstochowskiej
Instytut Inteligentnych Systemów Informatycznych
Wydział Inżynierii Mechanicznej i Informatyki
Politechnika Częstochowska
al. Armii Krajowej 36
42-200 Częstochowa

Recenzja

rozprawy doktorskiej mgra inż. Karola Grzegorczyka, pt.: Vector representations of text data in deep learning.

Promotor: prof. dr hab. inż. Witold Dzwinel

Promotor pomocniczy: dr inż. Marcin Kurdziel

Niniejszą recenzję opracowano na zlecenie Dziekana Wydział Informatyki, Elektroniki i Telekomunikacji AGH, Prof. dr hab. inż. Krzysztof Boryczko, z dnia 06.07.2018 r.

1. Charakterystyka tematu, celu i tezy badawczej rozprawy

Analiza danych tekstowych i przetwarzanie języka naturalnego jest ważnym obszarem eksploracji danych i uczenia maszynowego. Obecny świat produkuje ogromne ilości informacji we wszelkich możliwych formatach, co powoduje konieczność powstawania coraz bardziej wyszukanych form jej reprezentacji. W przypadku danych tekstowych od kilkudziesięciu lat trwają prace nad metodami uproszczonej reprezentacji za pomocą wektorów liczb. Dzięki takiej reprezentacji jest możliwa klasyfikacja lub wyszukiwanie podobnych tekstów za pomocą odpowiednich algorytmów. Cały czas jednak istnieje potrzeba udoskonalania istniejących metod reprezentacji tekstu. Autor deklaruje w pracy stworzenie wysokiej jakości gęstej reprezentacji dokumentów nadającej się do przetwarzania przybliżonymi metodami najbliższych sąsiadów oraz całościowego różniczkowalnego modelu osadzania słów wieloznacznych.

2. Zawartość rozprawy

Recenzowana praca mgra inż. Karola Grzegorczyka składa się z pięciu rozdziałów, spisu tabel i rysunków, bibliografii oraz dodatków. Dokument liczy 99 stron.

Rozdział pierwszy zawiera wprowadzenie do uczenia maszynowego, podział ze względu na rodzaj zadań oraz podstawy uczenia na podstawie danych. Pokrótce przedstawia problemy

związane z wektorowym opisem danych tekstowych, a mianowicie brak informacji o kolejności słów w modelu bag of words czy nierozróżnianie słów wieloznacznych lub homonimów w przypadku osadzania słów. Zdefiniowany jest również ogólny cel pracy: stworzenie wysokiej jakości gęstej reprezentacji dokumentów nadającej się do przetwarzania przybliżonymi metodami najbliższych sąsiadów oraz całościowego różniczkowalnego modelu osadzania słów wieloznacznych.

Rozdział 2 zawiera wybrane zagadnienia związane z uczeniem maszynowym. Przedstawiono ogólny model systemu uczącego się, podzielono uczenie na nadzorowane i nie oraz według zadań. Opisano ogólna zasadę uczenia z danych. Następnie przedstawia wektorowe sposoby reprezentacji danych, od najprostszej VSM, poprzez bag-of-words, modelowanie tematyczne, aż do metod uwzgledniających znaczenie słów, t.j. radzenia sobie z dokumentami, w których występują słowa wieloznaczne i homonimy. Następnie omówiono reprezentacje oparte na sieciach neuronowych poczynając od ogólnego opisu sieci jednokierunkowych i ich uczenia. Ponadto zebrano informacje o modelach neuronowych uczących się związków pomiędzy zestawami słów w obszarach znaczeniowych. Następnie omówiono metody głębokiego uczenia oraz grupowania danych. Rozdział kończy wykres przedstawiający reprezentacje wektorowe dokumentów tekstowych w przestrzeni poziomu reprezentacji oraz gęstości. Osie wykresu nie są opisane, ale wielkości reprezentowane przez nie są raczej oczywiste.

Rozdział 3 dotyczy reprezentacji na poziomie rozdziałów i zawiera dwa autorskie podejścia do tworzenia kodów binarnych o małej liczbie wymiarów. Autor proponuje tworzenie hashy na bazie całych dokumentów, w odróżnieniu od ich tworzenia na podstawie wygenerowanej wcześniej reprezentacji wektorowej. W tym celu stosuje małą sieć neuronową z tzw. zaokrągloną sigmoidą (*rounded logistic function*). Porównał również różne warianty binarnych stochastycznych neuronów i wybrał najlepszy do celów binaryzacji. Model został zaimplementowany w bibliotece TensorFlow. Autor w proponowanej metodzie wykorzystał do generowania osadzeń funkcję *embedding_lookup* biblioteki TensorFlow.

W Rozdziale 4 Autor przedstawia autorski model wieloznacznego osadzania słów. Rozwiązanie jest rozszerzeniem modelu skip-gram poprzez uczenie wektorów słów wieloznacznych aby przewidywać sąsiednie słowa w kontekście. Autor zaproponował nowy sposób regulacji uczenia nazwany *parallel penalty*, który nagradza brak równoległości wektorów wag. W jego ramach opracował trzy rodzaje współczynnika nieprostopadłości i wybrał eksperymentalnie jeden z nich.

Rozdział 5 przedstawia wnioski i sugeruje dużą liczbę kierunków dalszych badań.

Dodatek A zawiera dwa nowe usprawnienia metody bag of words. Autor zauważył, że metodach opartych na BoW wszystkie znaczenia słów wieloznacznych lub homonimów maja ta sama reprezentację. Metody, uwzględniające znaczenie wymagają dużych korpusów. Autor zmodyfikował metodę *multi-prototype vector-space model*, aby mogła pracować z mniejszymi zbiorami danych, nazywając nowy algorytm *bag-of-senses*. Dalej Autor zaproponował rozwiązanie problemu z ważeniem słów przez częstość ich występowania, gdzie musimy przejść do reprezentacji wektorów liczb rzeczywistych. Proponuje nowy sposób uwzględniania częstości słów w domenie liczb całkowitych.

Dodatek A zawiera ponadto zastosowanie osadzania słów do profilowania autorów blogów internetowych.

Dodatek B zawiera opis oprogramowania do przeprowadzania eksperymentów z uczeniem maszynowym, począwszy od przeglądu popularnych bibliotek, poprzez dokładniejszy opis TensorFlow, aż do biblioteki głębokiego uczenia AGH. Następnie opisano zbiory danych użyte w pracy. Należy podkreślić, że Autor wybrał nowoczesne i duże zbiory danych odzwierciedlające dobrze poziom skomplikowania w rzeczywistym świecie.

Pracę kończy rozbudowana bibliografia składająca się ze stu kilkudziesięciu aktualnych pozycji.

Ogólnie zasadnicze i oryginalne rezultaty pracy można podsumować następująco:

- Opracowanie wyczerpującego wprowadzenia do tematyki wektorowej reprezentacji danych tekstowych oraz przeglądu literaturowego.
- Stworzenie nowego modelu reprezentującego tekst na poziomie słów na bazie sieci neuronowej.
- Stworzenie nowego modelu reprezentującego tekst na poziomie dokumentu na bazie sieci neuronowej.
- Opracowanie usprawnień metody bag of words uwzględniające słowa wieloznaczne i homonimy dla małych korpusów oraz częstość występowania słów w domenie liczb całkowitych.
- Zastosowanie osadzania słów do profilowania autorów blogów internetowych.
- Opracowanie przeglądu oprogramowania do głębokiego uczenia maszynowego.

Wymienione oryginalne metody przedstawione w pracy zostały opublikowane w kilku artykułach naukowych, głównie w materiałach konferencyjnych, w tym bardzo prestiżowej konferencji Proceedings of the 2nd Workshop on Representation Learning for NLP oraz jednym opublikowanym w czasopiśmie z Listy Ministerialnej A. Zaprezentowany materiał pokazuje, że Doktorant zrealizował cel pracy.

3. Uwagi krytyczne i wskazówki dotyczące rozprawy

Praca napisana jest schludnie i przejrzyście. Należy podkreślić fakt napisania pracy w języku angielskim Praca obfituje w czytelne rysunki oraz schematy. Poniżej zamieszczam kilka uwag i pytań. Uwagi te nie umniejszają wartości naukowej rozważanej rozprawy doktorskiej.

Strona 36: However, sometimes computations are slowed down by I/O operations. To speed up training we feed mini-batches to TensorFlow Global shuffling can be easily done using Apache Hadoop or Apache Spark distributed processing frameworks session from multiple threads. This way the overall training time is shortened approximately 6 times.

Dlaczego sześć razy? Czy nie jest to związane z danym systemem komputerowym?

Część zaprezentowanych metod operuje na mniejszej reprezentacji danych. Czy jest możliwe wyliczenie lub podanie zysku pamięci, a w przypadku wszystkich metod, porównania czasów lub złożoności obliczeniowej?

Czy jest możliwe porównanie części metod z reprezentacją typu „one-hot", ale na poziomie znaków (np. Xiang Zhang, Junbo Zhao, Yann LeCun, Character-level Convolutional Networks for Text Classification, NIPS 2015)?

Praca posiada kilka drobnych błędów, tzw. literówek, np. represtation, na str. 32

Brakuje czasami łączników między wyrazami:

Neural network based text representations -> Neural network-based text representations

Część zmiennych w tekście nie jest pisana kursywą, np.

„…where N is a total number of training examples and M is…" na str. 20,

Rys. 2.1, Rys. 2.3,

n-dimensional na stronie 10,

„k distinct groups", „k clusters" na stronie 31.

Rys. 2.6 posiada nieopisane osie.

4. Wnioski końcowe recenzji

Podsumowując recenzję stwierdzam, że Pan mgr inż. Karol Grzegorczyk w rozprawie doktorskiej „Vector representations of text data in deep learning":

- Zrealizował cel rozprawy,
- Uzyskał oryginalne rezultaty naukowe dotyczące reprezentacji dokumentów tekstowych dla potrzeb głębokiego uczenia maszynowego,
- Dokonał ciekawego wprowadzenia do tematyki,
- Stworzył przegląd wybranych bieżących pozycji literaturowych dotyczących tematu,
- Stworzył nowe modele reprezentujące tekst na poziomie słów oraz całego dokumentu na bazie sieci neuronowej.
- Usprawnił metodę bag of words uwzględniając słowa wieloznaczne i homonimy dla małych korpusów oraz częstość występowania słów w domenie liczb całkowitych.
- Zastosował osadzanie słów do profilowania autorów blogów internetowych.
- Wykazał się umiejętnością samodzielnej pracy badawczej, znajomością literatury światowej i wiedzą w dziedzinie uczenia maszynowego.

Recenzowana praca spełnia wymagania ustawy o tytule i stopniach naukowych w dyscyplinie naukowej Informatyka. Wnoszę o jej przyjęcie i dopuszczenie do publicznej obrony. Jednocześnie ze względu na wysoki poziom naukowy rozprawy, opublikowanie jednego artykułu w czasopiśmie z ministerialnej listy A oraz jednego w materiałach bardzo prestiżowej konferencji, wnioskuję o wyróżnienie rozprawy.

Robert Schaefer