УДК 004.8:004.032.26 DOI: https://doi.org/10.30837/IYF.IIS.2024.047

АНАЛІЗ РЕКУРЕНТНИХ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ, ЇХ СТРУКТУРА

Греков О. О.
Науковий керівник – д.т.н., проф. Бодянський Є. В.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна
e-mail: oleksandr.hrekov@nure.ua

This conference paper provides insights into Recurrent Neural Networks (RNNs) in various domains like language modeling and speech recognition. It discusses key concepts such as «Backpropagation Through Time» and «Long Short-Term Memory Units» essential for understanding RNNs. Recent advancements like «Attention Mechanism» and «Pointer Networks» are also explored, showcasing improved performance in RNN-based techniques. Challenges like vanishing gradients are addressed, and solutions like Deep Recurrent Neural Networks (DRNNs) and Bidirectional Recurrent Neural Networks (BRNNs) are discussed. The Encoder-Decoder architecture, exemplified by Sequence to Sequence (seq2seq) models, is examined, and Pointer Networks (Ptr-Nets) are introduced as effective solutions for combinatorial optimization problems.

У сфері машинного навчання, що швидко розвивається, рекурентні нейронні мережі (RNN) відіграють життєво важливу роль як основний інструмент у різних сферах, включаючи моделювання мови, розпізнавання мовлення, створення описів зображень та тегування відео. Такі мережі також мають досить гарну перспективу у майбутньому використання в різноманітних сферах.

Рекурентні нейронні мережі – це спеціалізовані архітектури нейронних мереж, призначені для аналізу послідовних даних, таких як текст, геноми або часові ряди. На відміну від нейронних мереж прямого поширення (MLP), RNN включають цикли, що дозволяє їм зберігати пам'ять про попередні вхідні дані та враховувати контекст за межами поточного входу. У свою чергу, разом із рекурентними нейронними мережами застосовуються такі ключові поняття, як «поширення в часі» і «одиниці довготривалої короткочасної пам'яті», які є важливими для розуміння того, як ШНМ навчаються і зберігають інформацію в часі; «механізм уваги» і «мережі вказівників», які демонструють передові технології, що підвищують продуктивність ШНМ в різних завданнях.

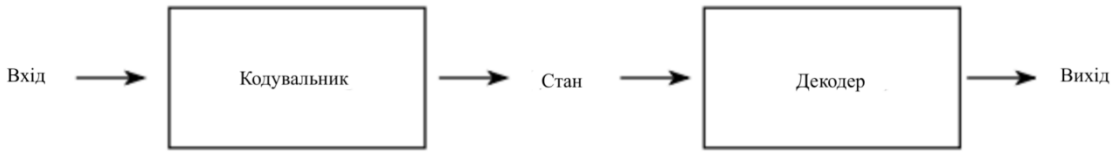
Поширення в часі (Backpropagation Through Time, BPTT) – це метод, заснований на алгоритмі зворотного поширення, спеціально адаптований для рекурентних нейронних мереж (RNN). Він перетворює RNN на традиційну нейронну мережу прямого поширення, дозволяючи зворотне поширення для оновлення ваг. Цей процес обчислює приховані та вихідні

стани крок за кроком під час прямого проходу. Потім визначається функція втрат для вимірювання розбіжності між вихідними та цільовими значеннями, агрегованими за кроками оновлення. Для оновлення ваг у BPTT обчислюються часткові похідні по кожній ваговій матриці з використанням ланцюгового правила, подібно до стандартного зворотного розповсюдження. Це дозволяє коригувати ваги на основі накопичених помилок на різних часових кроках, оптимізуючи параметри мережі для мінімізації загальної функції втрат.

Зникаючі або вибухові градієнти створюють значні проблеми при навчанні рекурентних нейронних мереж (RNN), як і в багатьох інших нейромережевих архітектурах. LSTM вирішують цю проблему шляхом включення вентильних комірок, які зберігають інформацію поза звичайним потоком нейронної мережі. Ці комірки використовують вентилі та комірки пам'яті для керування потоком інформації та ефективного пом'якшення зникнення градієнта. Повна структура LSTM інтегрує ці компоненти, щоб забезпечити ефективне збереження та передачу інформації в мережі, долаючи обмеження, пов'язані зі зникненням градієнтів у традиційних RNN. Також слід зазначити, що що LSTM також є наразі однією з найпоширеніших архітектур для рекурентних нейронних мереж яка використовується в багатьох проектах, таких як Google Tesseract та ін.

Глибокі рекурентні нейронні мережі (DRNN) складаються з декількох звичайних шарів RNN для створення глибокої архітектури. Кожен шар передає свій прихований стан наступному шару, полегшуючи потік інформації через мережу. Вихід обчислюється з використанням прихованого стану останнього шару. Двонаправлені рекурентні нейронні мережі (BRNN) включають як прямі, так і зворотні приховані стани, щоб захопити контекст як з минулих, так і з майбутніх послідовностей. Це дозволяє краще виконувати завдання, що вимагають властивостей передбачення, такі як заповнення пропусків у реченнях.

Архітектура кодера-декодера – це основна структура нейронної мережі, що складається з кодера та декодера. Кодер перетворює вхідні дані у представлення стану, як правило, вектор або тензор, тоді як декодер реконструює цей стан у вихідні дані.



Ця архітектура є основою для таких моделей, як Sequence to Sequence (seq2seq), що використовуються в основному в таких додатках, як Google Translate і пристроях з голосовим управлінням.

У seq2seq і кодер, і декодер використовують рекурентні нейронні мережі (RNN), а прихований стан кодера передається декодеру. Кодер містить блоки RNN, які послідовно обробляють вхідні елементи. Ці штучні нейронні мережі, часто LSTM або GRU, підвищують продуктивність моделі. Вектор кодера, що представляє кінцевий прихований стан кодера, консолідує інформацію з попередніх входів, слугуючи початковим станом для самого декодера.

Декодер, що також складається з блоків ШНМ, прогнозує вихідні дані на кожному часовому кроці на основі попереднього стану. Цей ітеративний процес генерує вихідну послідовність, кожен елемент якої визначається поточним станом декодера. Загалом, архітектура кодера-декодера, прикладом якої є seq2seq, пропонує гнучкий фреймворк для різноманітних завдань, пов'язаних з послідовністю, що сприяє ефективній передачі інформації та прогнозуванню.

Мережі вказівників (Ptr-мережі) покращують модель seq2seq за допомогою уваги, відходячи від фіксованих вихідних категорій. Замість того, щоб генерувати вихідну послідовність безпосередньо, Ptr-мережі генерують серію вказівників, що вказують на елементи вхідної послідовності.

На практиці Ptr-мережі використовують адитивну увагу, для обчислення вихідних умовних ймовірностей шляхом оцінки релевантності між станами. Ця оцінка нормалізується за допомогою функції softmax, що забезпечує імовірнісну інтерпретацію результату. Ptr-мережі являють собою значний прогрес у моделях «від послідовності до послідовності», пропонуючи універсальне рішення для проблем, що вимагають динамічних категорій виходів.

Список використаних джерел:

1. Nakamoto P. Neural Networks and Deep Learning: Neural Networks & Deep Learning, Deep Learning, Blockchain Blueprint. Createspace Independent Publishing Platform, 2018. 152 p.

2. Бодянский Е. В., Руденко О. Г. Искусственные нейронные сети: архитектуры, обучение, применения. Харьков: ТЕЛЕТЕХ, 2004. 369.

3. Шафроненко А. Ю., Бодянський Є. В., Руденко Д. О. Модифікований рекурентний метод достовірної нечіткої кластеризації з використанням оптимізаційної процедури на основі косяків риб. Системи обробки інформації. 2023. № 1 (172). С. 92–96.

4. Schmidt R. M. Recurrent Neural Networks (RNNs): A gentle Introduction and Overview. 2019. URL: https://arxiv.org/pdf/1912.05911.pdf (Дата запиту: 07.03.2024).

5. Salem F. M. Recurrent Neural Networks. Cham : Springer International Publishing, 2022. URL: https://doi.org/10.1007/978-3-030-89929-5 (Дата запиту: 08.03.2024).