DOI 10.37882/2223–2966.2021.03.36

ОБЗОР МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ НЕЙРОННОГО МАШИННОГО ПЕРЕВОДА

OVERVIEW OF DEEP LEARNING METHODS FOR NEURAL MACHINE TRANSLATION

Chzhun Zhujyuj

Summary. The article discusses the main methods of neurolinguistic training. Three stages of machine translation development are revealed: rule-based, statistical, neural learning. Two popular methods of neural machine translation are described: recurrent and convolutional. Advanced neural machine learning methods are characterized: ConvS2S, Transformer, RNMT +. A forecast was made for the further development of neural text translation systems.

Keywords: machine translation, neural learning, neural learning methods, neurolinguistics, ConvS2S, Transformer, RNMT +.

Чжун Жуйюй
Аспирант, Национальный исследовательский университет Московский институт электронной техники
zry1988510@126.com

Аннотация. В статье рассмотрены основные методы нейролингвистического обучения. Раскрыто 3 этапа развития машинного перевода: на основе правил, статистический, нейронного обучения. Описано 2 популярных способа нейронного машинного перевода: рекуррентный и сверточный. Охарактеризованы продвинутые методы нейронного машинного обучения: ConvS2S, Transformer, RNMT+. Сделан прогноз по дальнейшему развитию систем нейронного перевода текста.

Ключевые слова: машинный перевод, нейронное обучение, методы нейронного обучения, нейролингвистика, ConvS2S, Transformer, RNMT+.

Машинный перевод (Machine Translation) (MT) — это классическое направление в NLP, которое исследует, как использовать компьютерное программное обеспечение для перевода текста или речи с одного языка на другой без участия человека. Поскольку задача MT имеет аналогичную цель с конечной целью NLP и AI, т.е. полное понимание человеческого текста (речи) на смысловом уровне, ему было уделено большое внимание в прошлом.

Задача машинного перевода имеет долгую историю исследований, и за последние десятилетия было предложено множество эффективных методов. Недавно, с развитием глубокого обучения, появился новый метод под названием нейронный машинный перевод (NMT). Он способен фиксировать длительную зависимость в предложении, что указывает на огромный потенциал в превращении в новую тенденцию мейнстрима. После примитивной первоначальной модели предлагается множество моделей NMT, некоторые из которых достигли больших успехов с самыми современными результатами.

История машинного перевода

Машинный перевод (MT) имеет долгую историю; происхождение этого направления можно проследить до 17 века. В 1629 году Rene Descartes придумал универсальный язык, который выражал одно и то же значение на разных языках и разделял один символ.

Основные этапы его развития: машинный перевод на основе правил, статистический машинный перевод и нейронный машинный перевод [1].

1. Машинный перевод на основе правил

Машинный перевод на основе правил — это первая разработка в машинном переводе, основанная на гипотезе, что все разные языки имеют свой символ, представляющий одно и то же значение. Потому что обычно слово на одном языке может найти соответствующее слово на другом языке с тем же значением.

В этом методе процесс перевода можно рассматривать как замену слова в исходном предложении. С точки зрения «основанного на правилах», поскольку разные языки могут представлять одно и то же значение предложения в разном порядке слов, метод замены слов должен основываться на правилах синтаксиса обоих двух языков. Таким образом, каждое слово в исходном предложении должно занимать соответствующую позицию в целевом языке.

Самый серьезный недостаток метода, основанного на правилах, заключается в том, что он игнорирует по-

требность в контекстной информации в процессе перевода, что снижает надежность машинного перевода на основе правил.

2. Статистический машинный перевод

Статистический машинный перевод (SMT) был основной технологией в течение последних 20 лет. Он успешно применяется в отрасли, включая перевод Google, перевод Baidu.

Модель SMT находит слова (или фразы), которые имеют одинаковое значение, через двуязычный корпус по статистике.

Наиболее распространенной версией SMT является SMT на основе фраз (PBSMT), который, как правило, включает предварительную обработку, выравнивание предложений, выравнивание слов, извлечение фраз, подготовку функций фразы и обучение языковой модели. Ключевым компонентом модели PBSMT является лексика на основе фраз, в которой фразы на исходном языке сочетаются с фразами на целевом языке. Лексика построена из набора обучающих данных, который представляет собой двуязычный корпус. Таким образом, PBSMT может превзойти простые методы дословного перевода.

12. Нейронный машинный перевод

Из-за низкой производительности на начальном этапе и ограничений вычислительного оборудования соответствующие исследования по трансляции с помощью нейронной сети игнорировались в течение многих лет.

В связи с распространением глубокого обучения в 2010 году все больше и больше задач NLP достигли значительных улучшений. Большое внимание уделяется также использованию глубоких нейронных сетей для задач машинного перевода. Успешная модель нейронного машинного перевода (NMT) на основе DNN была впервые предложена Kalchbrenner and Blunsom, что к тому времени являлось совершенно новой концепцией для машинного перевода. По сравнению с другими моделями, модель NMT требует меньше лингвистических знаний, но может обеспечить конкурентоспособные результаты. С тех пор многие исследователи сообщили, что NMT может работать намного лучше, чем традиционная модель SMT, и она также широко применяется в промышленности.

Существует множество вариантов дизайна сети для NMT, которые можно разделить на повторяющиеся или неповторяющиеся модели. В частности, эту категорию можно проследить до раннего развития NMT, когда модели на основе RNN и CNN являются наиболее распространенным дизайном. Многие предложенные впоследствии сложные модели также принадлежат к семейству CNN или RNN. Этот подраздел следует за развитием NMT в первые годы и демонстрирует некоторые характерные модели, классифицируя их как модели на основе RNN или CNN.

1) NMT на основе RNN

Первый успешный NMT на основе RNN был предложен Sutskever и др., который использовал чистую модель глубокого RNN и получил производительность, которая приближается к лучшему результату, достигнутому SMT [2]. Дальнейшее развитие предложило механизм внимания, который значительно улучшает производительность перевода и превосходит лучшую модель SMT. Модель GNMT была отраслевой моделью, применяемой в Google Переводчике, и считалась важной вехой в NMT на основе RNN.

Помимо упомянутой выше работы, другие исследователи также предложили различные архитектуры с отличной производительностью. Zhang и др. предложили вариационный метод NMT, который имеет инновационные перспективы в моделировании задачи перевода, и соответствующий эксперимент показал лучшую производительность, чем исходный уровень оригинального NMT в задачах китайско-английского и англо-немецкого перевода. Чжоу и др. разработали Fast-Forward Connections для RNN (LSTM), которые могут позволить более глубокую сеть в реализации и, таким образом, получить лучшую производительность. Поскольку большее количество параметров часто означает лучшую репрезентативную способность, это демонстрирует огромный потенциал в будущем.

2) NMT на основе CNN (сверточной нейронной сети).

По сравнению с NMT на основе RNN, модели на основе CNN имеют преимущество в скорости обучения; это связано с внутренней структурой CNN, которая позволяет выполнять параллельные вычисления для различных фильтров при обработке входных данных. Кроме того, структура модели упростила модели на основе CNN для решения проблемы исчезновения градиента. Однако есть два фатальных недостатка, влияющих на качество их перевода. Во-первых, поскольку исходная модель на основе CNN может фиксировать зависимости слов только в пределах ширины своих фильтров, длинную зависимость слов можно найти только в высокоуровневых сверточных слоях. Этот неестественный характер часто вызывает худшую производительность, чем в моделях на основе RNN. Во-вторых, поскольку

исходная модель NMT сжимает предложение до фиксированного размера вектора, значительное снижение производительности произойдет, если предложение станет слишком длинным. Это происходит из-за ограниченной способности представления в фиксированном размере вектора. Подобное явление также можно найти в ранее предложенных моделях на основе RNN, которые позже смягчены механизмом внимания.

Также были предложены некоторые передовые модели NMT на основе CNN с соответствующими решениями для устранения вышеуказанных недостатков. Kaiser и др. предложил глубинно разделимые свертки на основе NMT. Созданный ими SliceNet может иметь аналогичную производительность с Kaiser и др. (2016). Gehring и др. (2017) последовали за своей предыдущей работой, предложив NMT на базе CNN, который сотрудничает с Attention Mechanism. Он даже получил лучший результат, чем модель на основе RNN, но вскоре это достижение было превзойдено Transformer [3].

Проведенные исследования предложили различные методы как в процессе обучения, так и в процессе вывода. Эти методы можно условно разделить на три категории в зависимости от их ориентации.

Первый интуитивно ориентирован на поиск решений для повышения скорости вычислений, которые могли бы поддерживать более обширный словарный запас. Второй фокусируется на использовании контекстной информации. Этот метод может адресовать некоторые из неизвестных слов (например, имя собственное), копируя их в результат перевода, а также редко встречающиеся слова, вызывающие плохое качество перевода.

Последний, более продвинутый, предпочитает использовать информацию внутри слова, такую как символы, из-за их гибкости в обработке морфологических вариантов слов. Этот метод может поддерживать перевод OOV-слов (сокращений) более «умным» способом.

1) Методы ускорения вычислений

Первая мысль при попытке увеличить скорость вычислений — это масштабировать операцию softmax. Поскольку эффективное вычисление softmax, очевидно, могло бы поддержать больший словарный запас, подобным попыткам уделялось много внимания в литературе по NLM. Morin и Bengio [4] предложили иерархические модели для экспоненциального ускорения вычисления коэффициента нормализации, таким образом помогая ускорить градиентный расчет вероятностей слов. В частности, оригинальная модель преобразовала словарь в двоичную древовидную структуру, которая была построена с предварительными знаниями из WordNet.

Первоначальный результат эксперимента показывает, что этот иерархический метод сопоставим с традиционной триграммной LM, но не может превзойти исходный NLM; отчасти это связано с использованием ручной работы из WordNet в процессе построения дерева.

Более элегантный метод — сохранить исходную модель, но изменить метод расчета коэффициента нормализации. Bengio и Senecal предложили метод выборки по важности для аппроксимации коэффициента нормализации. Однако этот метод не является стабильным без тщательного контроля. Mnih & Teh использовали шум-контрастную оценку для непосредственного изучения коэффициента нормализации, который может быть более стабильным в процессе обучения NLM.

Общей слабостью всех этих методов является то, что они по-прежнему страдают от слов OOV, несмотря на больший размер словаря, который они могут поддерживать. Это связано с тем, что расширенный словарь по-прежнему ограничен по размеру, и нет решения для дополнения при обнаружении неизвестных слов, тогда как следующая категория методов может частично справиться с этим. Кроме того, простое увеличение словарного запаса может просто немного улучшить из-за закона Ципфа, что означает, что всегда есть большой хвост OOV-слов, требующих обработки.

2) Методы с использованием контекстной информации

Более продвинутая методика использует контекстную информацию. Луонг и др. предложил алгоритм выравнивания слов, который взаимодействует с механизмом копирования для последующей обработки результата перевода. Конкретно, в методе Луонга для каждого слова OOV существует «Указатель», который отображается на соответствующее слово в исходном предложении. На этапе постобработки предопределенный словарь был снабжен «указателем» для поиска соответствующего перевода, при этом использовался механизм прямого копирования для обработки OOV слов, которых нет в словаре.

Популярность метода Luong частично объясняется тем, что механизм копирования фактически предоставляет бесконечный словарный запас. Кроме того, он синтезировал механизм копирования с общей операцией трансляции, добавив так называемую сеть коммутации, чтобы решить, какая операция должна применяться на каждом временном шаге [5]. Gu и др.

приложив параллельные усилия по интеграции различных механизмов, предложили своего рода механизм внимания, называемый CopyNet, с ванильной моделью кодировщика-декодера, которую можно естественным образом расширить для обработки слов OOV в задаче NMT. Вдобавок они обнаружили, что механизм внимания больше зависит от семантики и языковой модели при использовании традиционного перевода слов, но от местоположения при использовании операции копирования.

Вкратце, было предложено множество контекстно-зависимых уточнений, большинство из которых использует механизм копирования для обработки слов OOV с различными алгоритмами выравнивания, чтобы найти соответствующее слово на целевой стороне. Однако у этих методов есть ограниченные возможности для дальнейшего улучшения, поскольку механизм копирования слишком груб для обработки сложных сценариев на разных языках. На практике эти методы плохо работают на языках с богатой морфологией, таких как финский и турецкий.

3) Методы с мелким зерном

Фокусируются на использовании дополнительной информации внутри словарной единицы. Совершенно очевидно, что такая дополнительная информация могла бы улучшить способность охватить различные языковые явления.

В предыдущих исследованиях, хотя использование семантической информации о словарной единице могло обеспечить подавляющее большинство возможностей обучения, другие особенности на уровне подслова обычно игнорировались. С лингвистической точки зрения понятие «слово» является базовой единицей языка, но не является минимальным по содержанию семантической информации, и существует множество опытных правил, которые можно узнать изнутри таких единиц слова, как форма и суффикс.

Один из популярных вариантов использования подслова, был предложен Sennrich и др., и, как было доказано, он имеет лучшую производительность в некоторых общих результатах. Конкретно, в этой методике неизвестные слова рассматриваются как последовательности подсловных единиц, что является разумным с точки зрения состава подавляющего большинства этих слов (например, именованных сущностей, заимствованных слов и морфологически сложных слов).

Однако восстановление подслов приводит к огромному расходу места в словарном запасе, что фактически сводит на нет цель снижения вычислительной эффективности как во времени, так и в пространстве. В этих условиях для операции сегментации слов на обеих сторонах языков был применен метод извлечения подслов на основе парного байтового кодирования (BPE), который успешно адаптировал этот старый, но эффективный метод сжатия данных при предварительной обработке текста.

Метод BPE развивался для получения лучшего обобщения. Таку предложил регуляризацию подслов в качестве альтернативы для обработки феномена ложной неоднозначности в BPE, а также они предложили новый алгоритм сегментации подслов, основанный на модели униграммы, которая разделяла концепцию BPE, но была более гибкой в получении нескольких сегментаций на основе их вероятностей. Аналогичным образом Wu и др. использовал концепцию «заготовок» для обработки слов OOV, которые когда-то применялись в системе распознавания речи Google для решения проблемы сегментации японского и корейского языков. Этот метод разбивает слова на части, чтобы получить баланс между гибкостью и эффективностью при использовании отдельных символов и целых слов по отдельности.

Продвинутые модели NMT

1. ConvS2S

ConvS2S — это сокращение от Convolutional Sequence to Sequence, которая представляет собой сквозную модель NMT, предложенную Gehring и др. [5]. В отличие от большинства моделей NMT на основе RNN, ConvS2S является полностью основанной на CNN моделью как в кодировщике, так и в декодере. В сетевой структуре ConvS2S объединила 15 уровней CNN в своем кодировщике и декодере с фиксированной шириной ядра 3. Эта глубокая структура помогает компенсировать слабые места в захвате контекстной информации (рис. 1).

Что касается деталей сети, ConvS2S применил стробированные линейные единицы (GLU) в построении сети, которые обеспечивают стробированную функцию для вывода сверточного слоя. В частности, выходной сигнал сверточного слоя $Y \in R^{2d}$, который представляет собой вектор с двойными размерами (2d числами измерений) встраивания каждого входного элемента (d числами измерений), стробированная функция обрабатывает выходной сигнал

Помимо нововведения в структуре кодера-декодера на основе CNN, ConvS2S также применил аналогичный механизм внимания, который был безоговорочно принят моделью RNN, под названием Multistep Attention. Конкретно, многоступенчатое внимание —

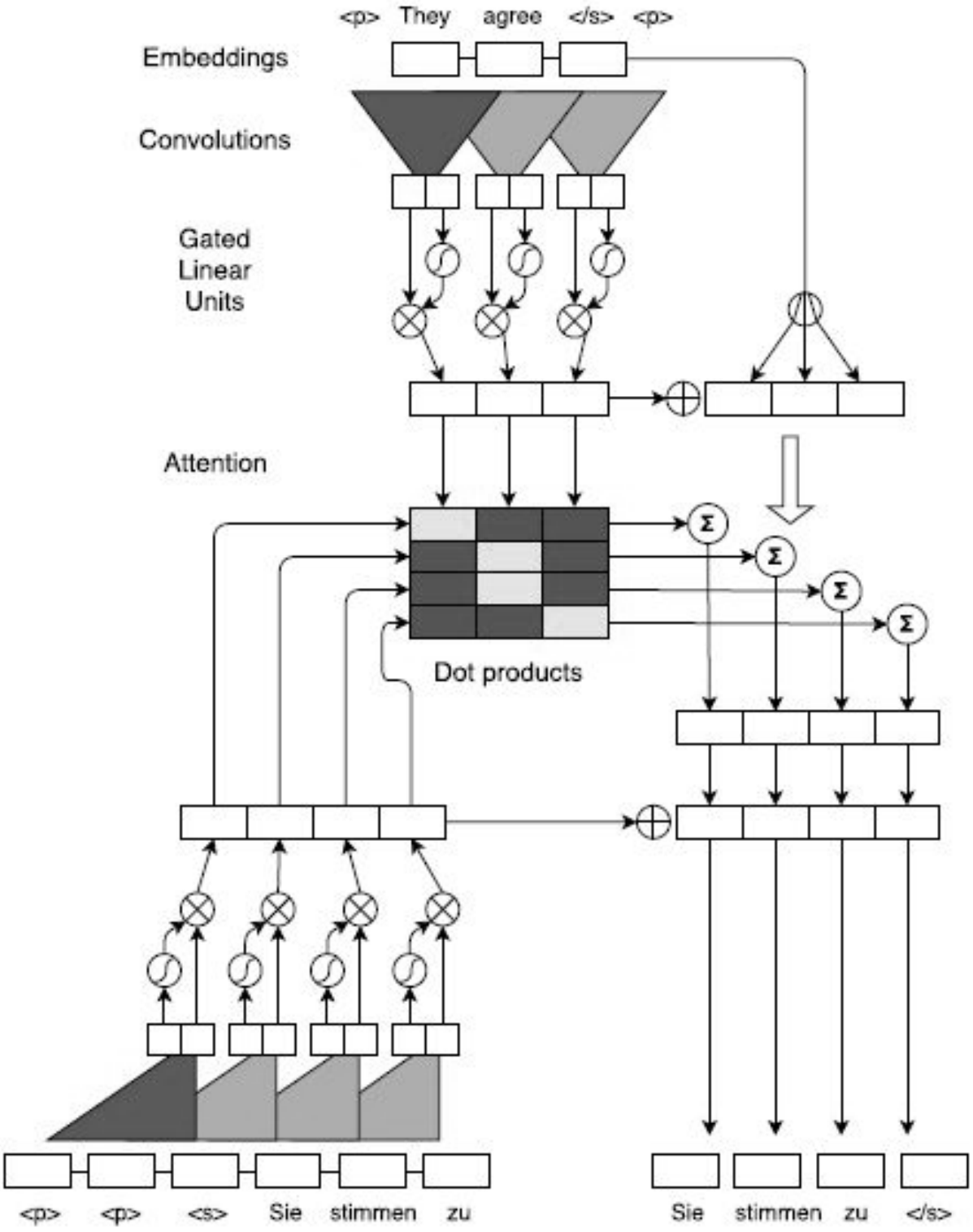

Рис. 1. Структура модели ConvS2S

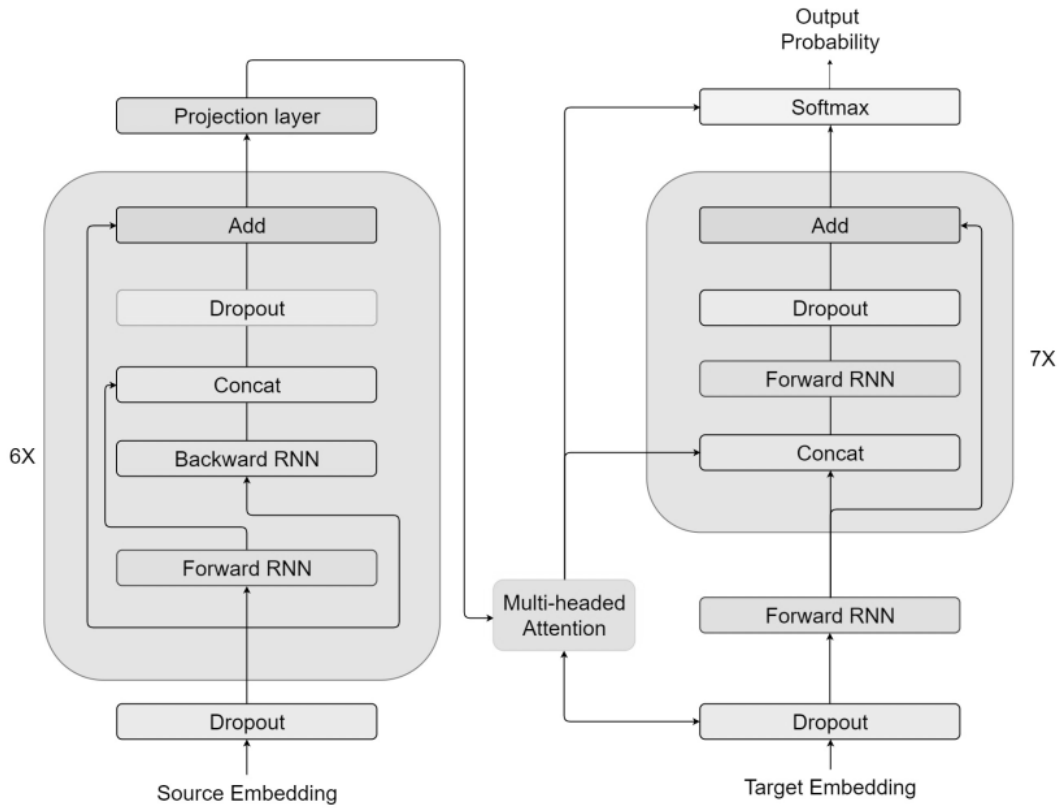

Рис. 2. Структура модели RNMT +, которая имеет аналогичную структуру GNMT с адаптивными инновациями в механизме внимания

это отдельная структура внимания, применяемая на каждом уровне декодера.

2. RNMT+

RNMT+ был предложен Chen и др. [6]. Эта модель напрямую унаследовала структуру модели GNMT, предложенную Wu и др. В частности, RNMT + можно рассматривать как улучшенную модель GNMT, которая продемонстрировала лучшую производительность модели NMT на основе RNN (рис. 2). По структуре модели RNMT + в основном отличается от модели GNMT в следующих нескольких аспектах.

Во-первых, RNMT+ использовал шесть двунаправленных RNN (LSTM) в своем декодере, тогда как GNMT использовал один уровень двунаправленной RNN с семью уровнями однонаправленных RNN. Эта структура принесла в жертву эффективность вычислений в обмен на исключительную производительность.

Во-вторых, стратегия синхронного обучения была предусмотрена в тренировочном процессе, улучшив скорость сходимости с производительность модели, основанной на эмпирических результатах.

3. Transformer и модели на его основе

Transformer — это новая структура NMT, предложенная Vaswani и др. [3]. В отличие от существующих моделей NMT, он отказался от стандартных структур RNN / CNN и разработал инновационные многослойные блоки самовнимания, которые объединены с методом позиционного кодирования. Эта новая тенденция проектирования структуры использует преимущества модели на основе как RNN, так и CNN, которая в дальнейшем использовалась для инициализации входного представления для других задач NLP. Примечательно, что Transformer — это полная модель NMT, основанная на внимании.

Transformer имеет свое уникальное представление при обработке входных данных, которое сильно отличается от рекуррентной или сверточной модели. Для вычисления самовнимания Transformer обрабатывает входные данные как три вида векторов для разных целей: это векторы ключа, значения и запроса. И все эти векторы управляются путем умножения содержимого входных данных на три матрицы, которые мы обучили в процессе обучения.

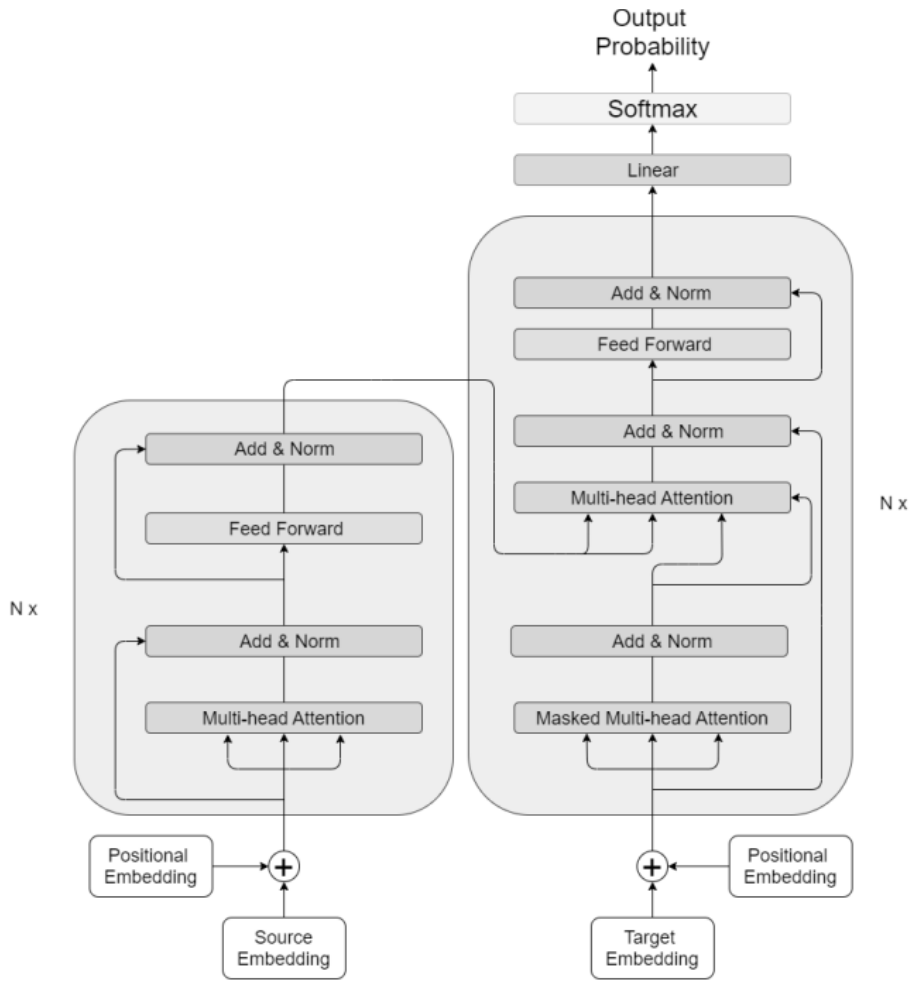

Рис. 3. Полная структура Transformer

Самовнимание — главное новшество в Transformer. Вместо того, чтобы просто вычислять самовнимание один раз, механизм с несколькими ядрами проходит через масштабированный скалярный продукт внимания несколько раз параллельно, а результаты этого независимого внимания затем объединяются и линейно преобразуются в ожидаемые измерения. Из-за огромного повышения производительности, произведенного Transformer, он привлек огромное внимание исследователей. Общепризнанный недостаток Transformer включает в себя: отсутствие рекуррентного моделирования, теоретически неполное по Тьюрингу, получение информации о местоположении, а также большую сложность модели. Все эти недостатки препятствовали дальнейшему повышению качества перевода. В ответ на эти проблемы были предложены некоторые корректировки.

Что касается архитектуры модели, то некоторые предложенные модификации были сосредоточены как на уровне глубокого внимания, так и на композиции сети. Варна и др. предложили преобразователь в 2–3 раза более глубокий с усовершенствованным механизмом внимания, который может быть проще для оптимизации более глубоких моделей.

В отличие от фиксированных слоев модели NMT, Dehghani и др. предложили Universal Transformers, которые отменили укладку постоянного количества слоев за счет объединения рекуррентного индуктивного смещения RNN и механизма остановки адаптивных вычислений, таким образом улучшив исходное представление на основе самовнимания для лучшего обучения итерационным или рекурсивным преобразованиям. Примечательно, что эта корректировка показала, что модель является полной по Тьюрингу при определенных предположениях.

Что касается уточнения композиции сети, вдохновленной идеей AutoML, So и др. применил поиск нейронной архитектуры (NAS), чтобы найти сопоставимую модель с упрощенной архитектурой. Evolved

Transformer имеет инновационную комбинацию базовых блоков и достигает того же качества, что и оригинальная модель Transformer-Big, с меньшими параметрами на 37,6%.

Хотя большая часть изменений направлена на непосредственное изменение структуры модели, в некоторых новых публикациях было решено использовать другое представление входных данных для повышения производительности модели. Один из прямых методов — это использование улучшенного кодирования положения для внедрения порядка следования. Shaw и др. предложили модифицированный механизм самовнимания с осознанием использования представлений об относительных положениях, который продемонстрировал значительные улучшения в двух задачах MT.

Одновременно с этим, использование предварительно инициализированного представления ввода с точной настройкой — это еще одно направление, где были предложены некоторые идеи в различных задачах NLP, таких как применение ELMo для кодировщика модели NMT. Что касается Transformer, одним из побочных продуктов этой инновационной модели является использование самовнимания для представления последовательности, которое может эффективно объединять словесную информацию с контекстной информацией.

Были предложены два хорошо известных метода представления входных данных на основе трансформера, названные Bert (представление двунаправленного кодера от Transformers) и GPT (Generative Pretrained Transformer), для улучшения некоторых последующих задач NLP.

Полная структура трансформатора проиллюстрирована на рис. 3.

Аспекты развития проблемы на перспективу

Хотя мы стали свидетелями быстро растущего прогресса исследований в области NMT, остается еще много проблем. Суммируем основные проблемы и перечисляем некоторые потенциальные направления в нескольких аспектах.

1. Что касается производительности перевода, NMT по-прежнему не справляется с переводом длинных предложений. Это происходит в основном из-за двух причин: практических ограничений в разработке и способности самой модели к обучению.
2. Механизм выравнивания важен как для моделей SMT, так и для моделей NMT. Мы считаем, что этот продвинутый метод выравнивания будет привлекательным для будущих исследований, поскольку мощный метод внимания может напрямую улучшить характеристики модели. Более поздние исследования механизма внимания попытаются ослабить слабость NMT, такую как способность интерпретации.
3. Низкоресурсный нейронный машинный перевод — еще одна горячая точка в нынешнем NMT, который пытается решить проблему серьезного снижения производительности, когда модель NMT обучается с редким двуязычным корпусом.
4. Наконец, исследования в области приложений NMT станут более обширными. В настоящее время разработано множество приложений, таких как перевод речи и перевод на уровне документа.

ЛИТЕРАТУРА

1. Cho K., Van Merrinboer, B., Bahdanau D., & Bengio Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches.arXiv preprint arXiv:1409.1259.
2. Sutskever I., Vinyals O., & Le Q.V. (2014). Sequence to sequence learning with neural networks. InAdvances in neural information processing systems(pp. 3104–3112).
3. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A N., … & Polosukhin I. (2017). Attention is all you need. InAdvances in Neural Information Processing Systems(pp. 5998–6008).
4. Morin F., & Bengio Y. (2005, January). Hierarchical probabilistic neural network language model. InAistats(Vol. 5, pp. 246–252).
5. Gulcehre C., Ahn S., Nallapati R., Zhou B. & Bengio Y. (2016). Pointing the unknown words. arXiv preprint arXiv:1603.08148.
6. Chen M. X., Firat O., Bapna A., Johnson M., Macherey W., Foster G., … & Wu Y. (2018). The best of both worlds: Combining recent advances in neural machine translation. arXiv preprint arXiv:1804.09849.
7. Clinchant S., Jung K.W., & Nikoulina V. (2019). On the use of BERT for Neural Machine Translation. arXiv preprint arXiv:1909.12744.

Журнал «Современная наука: актуальные проблемы теории и практики»