**Слайд 1**

Приветствие.

Тема доклада: Слой внимания в нейронных сетях

**Слайд 2**

Давайте рассмотрим пример, где нам нужно узнать человека по фотографии нескольких известных людей. В основном это групповая фотография людей, которых мы знаем. Теперь нам нужно распознать только одного человека, так как наш разум помогает нашему подсознанию. Ум создаст образ этого человека, и, сопоставив его, мы сможем узнать этого человека. Это означает, что наш разум обращает внимание только на образ этого человека, который был создан. Таким образом, сосредоточение внимания только на одном человеке в группе можно рассматривать как внимание.

До введения механизма внимания базовая модель LSTM или RNN была основана на системе кодер-декодер.

Механизм внимания — один из самых ценных прорывов в подготовке моделей глубокого обучения за последние несколько десятилетий. Он широко использовался в задачах обработки естественного языка (Natural Language Processing - NLP). Слой внимания может помочь нейронной сети запоминать большие последовательности данных.

**Слайд 3**

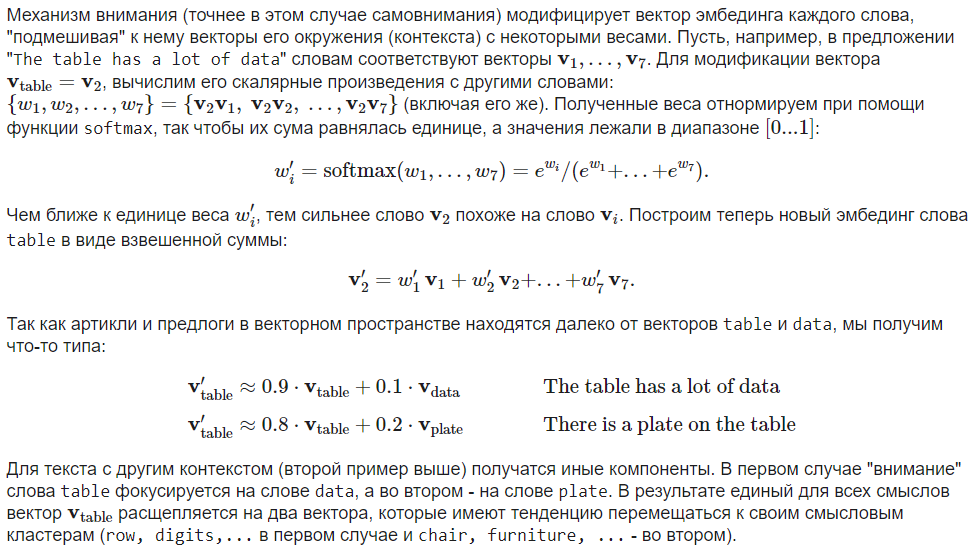
Одно слово по-разному «обращается» к другим словам в одном и том же предложении. Когда мы видим слово «eating», мы ожидаем, что очень скоро встретим, что конкретно, какой продукт.

Внимание в глубоком обучении можно широко интерпретировать как вектор весов важности: чтобы предсказать или сделать вывод об одном элементе, таком как пиксель в изображении или слово в предложении, мы оцениваем, используя вектор внимания, насколько сильно он коррелирует или «уделяет внимание» другим элементам и принимает сумму их значений, взвешенных по вектору внимания, в качестве аппроксимации цели.

**Слайд 4**

При работе с естественным языком, каждому слову в словаре ставится в соответствие вектор с вещественными компонентами (эмбединг слова). Эти векторы поступают на вход нейронной сети с той или иной архитектурой. Компоненты векторов являются параметрами, которые подбираются в процессе обучения так, чтобы близкие по смыслу слова имели схожие векторы. В качестве меры близости обычно выступает косинус угла между векторами, который определяется их скалярным произведением.

**Слайд 5**



**Слайд 6**

Механизм внимания был введен для повышения производительности модели кодер-декодер для машинного перевода. Идея механизма внимания заключалась в том, чтобы позволить декодеру гибко использовать наиболее релевантные части входной последовательности путем взвешенной комбинации всех закодированных входных векторов, при этом наиболее релевантным векторам присваивались самые высокие веса.

**Слайд 7**

Механизм внимания был представлен Bahdanau et al. (2014) для решения проблемы узкого места, возникающей при использовании вектора кодирования фиксированной длины, когда декодер будет иметь ограниченный доступ к информации, предоставленной на входе. Считается, что это становится особенно проблематичным для длинных и/или сложных последовательностей, где размерность их представления должна быть такой же, как и для более коротких или простых последовательностей.

Обратите внимание, что механизм внимания Бахданау и др. разделен на пошаговые вычисления оценок выравнивания, весов и вектора контекста :

Показатели выравнивания: модель выравнивания принимает закодированные скрытые состояния, hi, и предыдущий вывод декодера, st-1, чтобы вычислить балл, et,i, который указывает, насколько хорошо элементы входной последовательности совпадают с текущим выходом в позиции, t. Модель выравнивания представлена ​​функцией, a(.), который может быть реализован нейронной сетью с прямой связью:

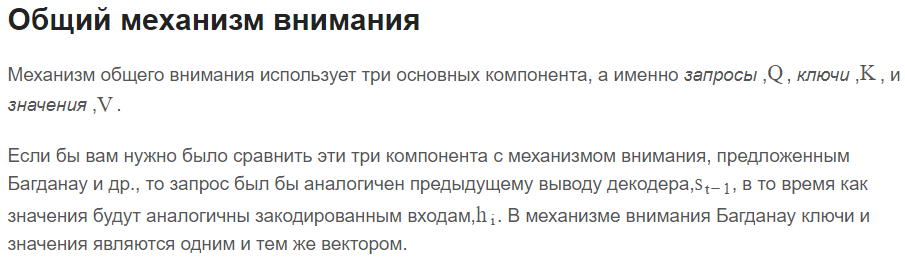
Веса: веса, at,i, вычисляются путем применения операции softmax к ранее вычисленным показателям выравнивания:

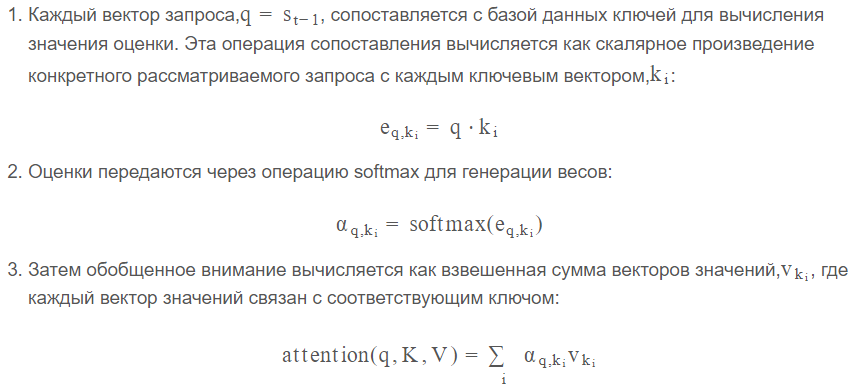
Вектор контекста: уникальный вектор контекста, ct, подается в декодер на каждом временном шаге. Он рассчитывается как взвешенная сумма всех, T, скрытые состояния кодировщика:

Багданова и др. реализовал RNN как для кодировщика, так и для декодера.

Однако механизм внимания может быть переформулирован в общую форму, которая может быть применена к любой задаче от последовательности к последовательности (сокращенно seq2seq), где информация не обязательно может быть связана последовательно**.**

**Слайд 8**





В контексте машинного перевода каждому слову во входном предложении будут присвоены собственные векторы запроса, ключа и значения. Эти векторы генерируются путем умножения представления кодера конкретного рассматриваемого слова на три разные матрицы весов, которые должны были быть сгенерированы во время обучения.

По сути, когда механизм обобщенного внимания представлен последовательностью слов, он берет вектор запроса, связанный с некоторым конкретным словом в последовательности, и оценивает его по каждому ключу в базе данных. При этом он фиксирует, как рассматриваемое слово связано с другими в последовательности. Затем он масштабирует значения в соответствии с весовыми коэффициентами внимания (вычисляемыми на основе оценок), чтобы удерживать внимание на тех словах, которые относятся к запросу. При этом он производит вывод внимания для рассматриваемого слова.

**Слайд 9**

Реализуем простую модель BI-LSTM и BI-LSTM with attention

**Слайд 10**

Загружаем библиотеки

**Слайд 11**

Импортируем набор данных.

В моделях используеся keras.dataset, предоставленный набором данных imdb.

Набор данных содержит классифицированные отзывы зрителей фильма.

**Слайд 12**

Определение модели Bi-LSTM.

**Слайд 13**

**Слайд 14**

Определение класса внимания.

**Слайд 15**

Определение класса внимания (продолжение).

**Слайд 16**

Создание модели с использованием слоя attention, где return\_sequence = true

**Слайд 17**

Мы видим, что точность и потери модели в данных изменились, мы получали точность около 94% для 12 эпох с использованием модели Bi-Lstm. После использования слоя «Attantion» в модели мы увеличили точность до 99%, а также снизили потери до 0,0285.

Механизм «внимания» повышает производительность модели и его можно использовать с любой моделью рекуррентных нейронных сетей.

**Слайд 18**

Механизм «внимания» повышает производительность модели и его можно использовать с любой моделью рекуррентных нейронных сетей.