Федеральное государственное автономное образовательное учреждение высшего образования "Национальный исследовательский университет "Высшая школа экономики"

Факультет компьютерных наук

Основная образовательная программа

Прикладная математика и информатика

КУРСОВАЯ РАБОТА

Исследовательский проект на тему "Оценка неопределенности для Машинного Перевода"

Выполнил студент группы 171, 3 курса: Кузнецов Дмитрий Сергеевич

Руководитель КР научный сотрудник Лобачева Екатерина Максимовна

Москва 2020

1 Аннотация

В области машинного перевода *beam search* - один из основных методов улучшения качества итогового предсказания. Однако при использовании нейросетевых технологий возникает проблема больших beam width, что ограничивает богатство beam search. Решение данной проблемы поможет значительно повысить качества предсказаний засчет рассмотрения бОльшего числа гипотез в процессе построения предсказаний. Множество факторов инфраструктуры модели могут влиять на возникновение beam problem. В этой работе мы исследуем beam problem и ее взаимосвязь с *model uncertainty*. Также изучим склонность моделей с оценкой неопределенности к beam problem.

2 Ключевые слова

machine translation, uncertainty estimation, beam search, beam problem, model uncertainty

3 Введение

Задача машинного перевода заключается в предоставлении по исходному предложению на языке-источнике предложения, эквивалентного по смыслу, на целевом языке. Пусть $e_1, \ldots, e_n$ последовательность векторого представления токенов входного предложения на языке-источнике. От модели машинного перевода требуется построить последовательность токенов $f_1, \ldots, f_n$ на целевом языке, причем на этом языке данная последовательность должна иметь исходный смысл.

Большинство современных моделей нейромашинного перевода следует парадигме Sequence to sequence (?), далее *seq2seq*. Seq2seq модели зачастую представляют из себя две рекурентные нейронные сети или группы сетей. Одна сеть, обрабатывающая входные токены, называется *encoder*, вторая сеть, строящая выходные токены, называется *decoder*. Задача encoder заключается в кодировании входной последовательности в некоторое внутреннее представление $h_1, \ldots, h_p$, которое подается на вход decoder в качестве инициализации перед построением выходного предсказания. Выход decoder множество векторов $\{(p_{i1}, \ldots, p_{id})\}_{i=1}^{m}$, где d - объем словаря целевого языка, m - максимально возможная длина предсказываемого предложения, p_{ij} - уверенность модели в том, что i-ый токен предсказываемого предложения будет j-ым токеном из словаря целевого языка. В итоге модель перевода строит "вероятностное распределение"в пространстве декартова произведения целевого словаря.

Наиболее простым и очевидным способом построения предсказания (итогового перевода) - это выбор на каждом шаге токена, в котором модель наиболее уверена. Если $s_1, \dots, s_m$ последовательность токенов гипотезы модели, тогда $s_i := argmax \{p_{ij}\}_{j=1}^{d}$. Однако не верно утверждение о том, что всегда для оптимального предсказания с точки зрения уверенности модели $s_1^*, \dots, s_m^*$ выполнено $s_1^* = argmax \{p_{1j}\}_{j=1}^{d}$. Данное явление связано и с beam problem, которую мы подробнее разберем в последующих главах.

В качестве более эффективного решения построения предсказания используют beam search. Его основная идея заключается в том, что вместо выбора наиболее вероятного токена на каждом шаге потребуем от модели хранить b наиболее вероятных префиксов предложения $s_{b1}, \dots, s_{bt}$. С помощью такого подхода мы выбираем не локально оптимальный токен, а в совокупности оптимальный префикс, что повысит шансы на построение грамотного предсказания.

Несложно понять, что с увеличением b (данный гиперпараметр называется beam width) должно расти и качество предсказания, т.к. растет число жадных переборов. На практике оказывается, что в нейромашинном переводе начиная с некоторого порога с увеличением beam width качество показательно падает. Иллюстрацию данного явления можно видеть на рис. 1.1., здесь используется модель, описанная в статье Jonas Gehring (2017).

Подобная проблема называется *beam problem* и поднималась в статье Somebody (Sometm). Далее мы подробнее разберемся в возможных причинах этого явления, пока заметим, что не исключена следующая ситуация. Пусть на некотором шаге k_1 beam search сохранил среди b локально оптимальных префикс $s_1, \dots, s_{k_1}$, который имеет наименьшую уверенность среди всех b префиксов. Пусть этот префикс построен так, что на последующих шагах найдется дополняющий его суффикс так, что итоговое предложение считается оптимальным переводом с точки зрения модели. Бывают случаи, когда такой перевод окажется некорректным. Рассмотрим такой пример. Пусть мы обучали модель перевода на корпусе сказок, исходный язык - русский, целевой язык - английский. Рассмотрим следующее предложение: *Жульничество на экзамене однажды может закончиться проблемами.* Пусть наша модель выдала в качестве перевода: *At the examination once upon a time.* Очевидно, что смысл перевода не соответствует исходному предложению, тем не менее это предложение оптимально с точки зрения нашей гипотетической модели. Beam problem может стать причиной таких ситуаций. Префикс *At the examination once* мог попасть при широком beam search в рассмотрение и иметь низкую степень уверенности. Однако в силу того, что мы обучались на большом корпусе сказок, где встречается много примеров предложений *Once upon a time...* наша модель приняла решение, что после *Once* вероятно

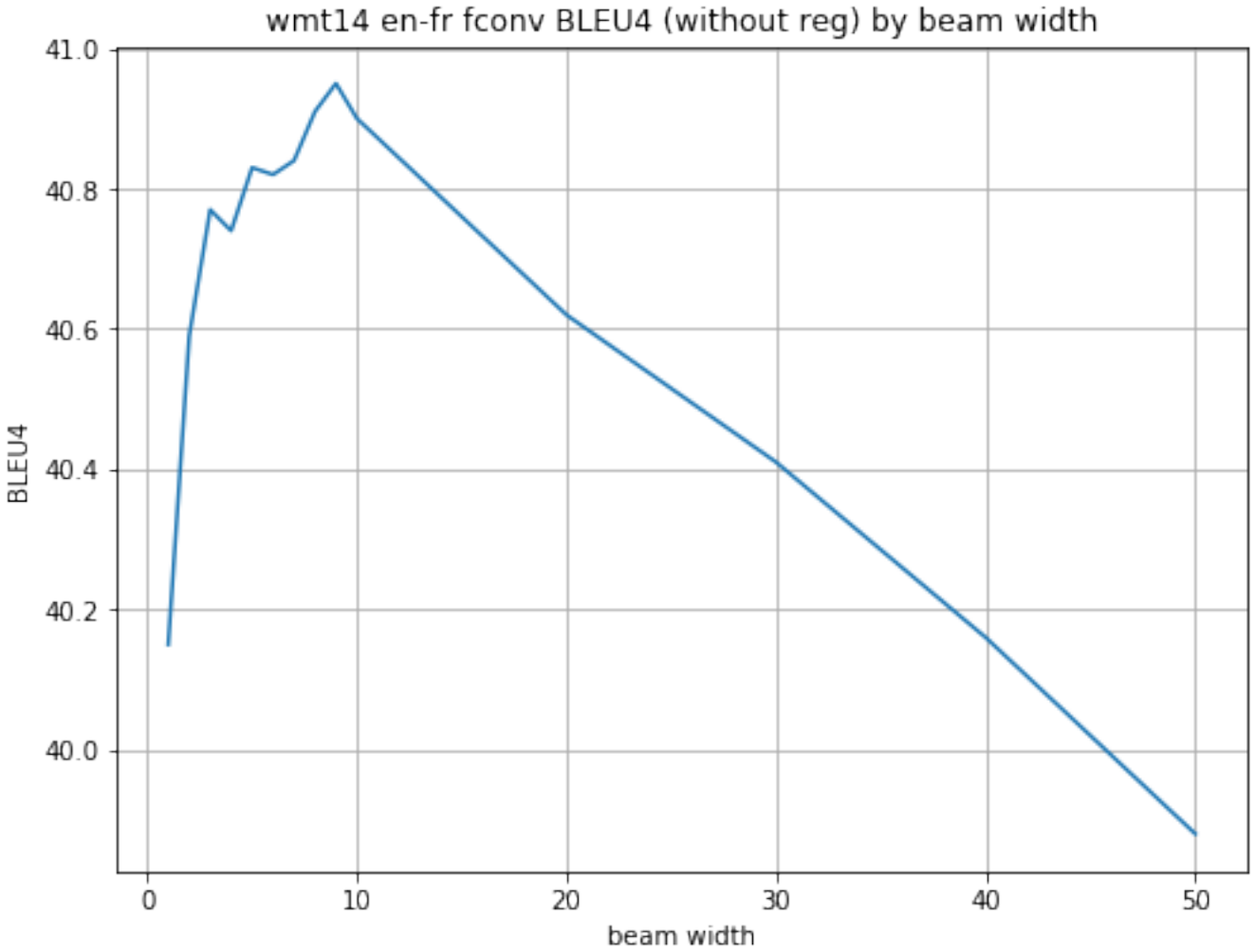

Рис. 1: Рис. 1.1. - Зависимость BLEU4 от beam width для Convolutional seq2seq, обученной на корпусе *WMT'14 English-France*

будет идти продолжение *upon a time*, что привело к плохом переводу.

Заметим, что в примере к плохому переводу привели некоторый особенности модели. Наша модель в качестве предсказания выдала некоторое мало вероятное в совокупности предсказания под действиям некоторых факторов. Подобные явления вызывают *uncertainty*, в частности *model uncertainty*, неопределенности предсказаний, вызванных моделью. Что если мы сможем построить некоторую вероятностную меру на пространстве выходных предложений $s_1, \ldots, s_m$. В таком случае мы сможем оценить в совокупности насколько свойственно получить то или иное предсказание и корректировать выход модели. Задачей явного или неявного построения подобного распределения на выходах называется *uncertainty estimation*.

Существуют статьи, в которых рассматриваются разные виды неопределенностей и их влияние на beam search. В данной работе мы исследуем влияние model uncertainty на beam problem. Также исследуем применимость методов uncertainty estimation в задаче

машинного перевода и их влияние на результативность beam search.

4 Beam Search

4.1 Определение метода

Положим префикс предсказания длины k:

$$f^i_{:k} := (f^i_1, \ldots, f^i_k), \quad \text{где i - номер векти beam search} \tag{1}$$

Введем следующий итеративный метод:

1 Подадим на вход decoder специальный токен начала предложения и скрытое состояние encoder. Получим некоторый вектор уверенностей в токенах для целевого словаря $(p^{00}_1, \ldots, p^{00}_d)$

2 Выберем $beam$ токенов, значения уверенностей которых являются $beam$ наибольшими среди $(p^{00}_1, \ldots, p^{00}_d)$. Здесь $beam$ - гиперпараметр метода и называется *beam width*. В итоге мы запомним: $f^1_{:1}, \ldots, f^{beam}_{:1}$.

3 На следующей итерации подадим decoder в качестве входов $(f^{00}_1, \ldots, f^{00}_d)$ и скрытое состояние предыдущего шага. Для каждого токена мы получим свое выходное распределение $(p^{1i}_1, \ldots, p^{1i}_d)$ (здесь 1 - номер итерации, i - номер токена).

4 Для всякого токена $i \in \overline{1, beam}$ рассмотрим $\forall j \in \overline{1, d}: \; p(f^i_{:1}) * p^{1i}_j$. Получим уверенности префиксов длины 2. Выберем среди всех таких $beam * d$ префиксов $beam$ с наибольшими уверенностями. Запомним их: $f^1_{:2}, \ldots, f^{beam}_{:2}$

5 Итеративно будем продолжать операцию $\forall k \in \overline{3, m}$. Будем получать: $f^1_{:k}, \ldots, f^{beam}_{:k}$

6 В качестве итогового предсказания выберем $f^i_{:m}, \ldots, f^i_{:m}$ такой, что его уверенность наибольшая $\forall i \in \overline{1, beam}$

Резюмируя, данный метод на каждой итерации декодирования хранит $beam$ "наиболее вероятных"префиксов, в отличие от наивного подхода, который жадно выбирает 1 оптимальную гипотезу на каждой итерации.

4.2 Beam Problem

Список литературы

1. Первая статья в которой поднималась проблема beam search

2. Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin. Facebook AI Research. Convolutional Sequence to Sequence Learning. 25 Jul 2017

3. Здесь должна быть статья по seq2seq

4. Uri Alon, Shaked Brody, Omer Levy, Eran Yahav. Code2Seq: Generating sequences from structured representations of code. In ICLR 2019.