Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks

Александра Рябинина

12 марта 2018

Задача

Задача генерации последовательности токенов произвольной длины

- Image captioning problem
- Constituency Parsing
- Speech Recognition

Задача

- Обучающая выборка: $\{X^i, Y^i\}_{i=1}^N$
- $\log P(Y|X) = \sum_{t=1}^{T} \log P(y_t|y_{t-1}, X) = \sum_{t=1}^{T} \log P(y_t|h_t, \theta)$
- $h_t = \begin{cases} f(X, \theta), & \text{если } t = 1 \\ f(h_{t-1}, y_{t-1}, \theta), & \text{иначе} \end{cases}$

Проблема

- Во время обучения модель использует реальные данные
- Во время генерации модель использует синтетические данные, сгенерированные ей самой
- Модель склонна к накоплению ошибки: ранняя ошибка в генерации последовательности используется в качестве входа для генерации следующего токена

Scheduled Sampling

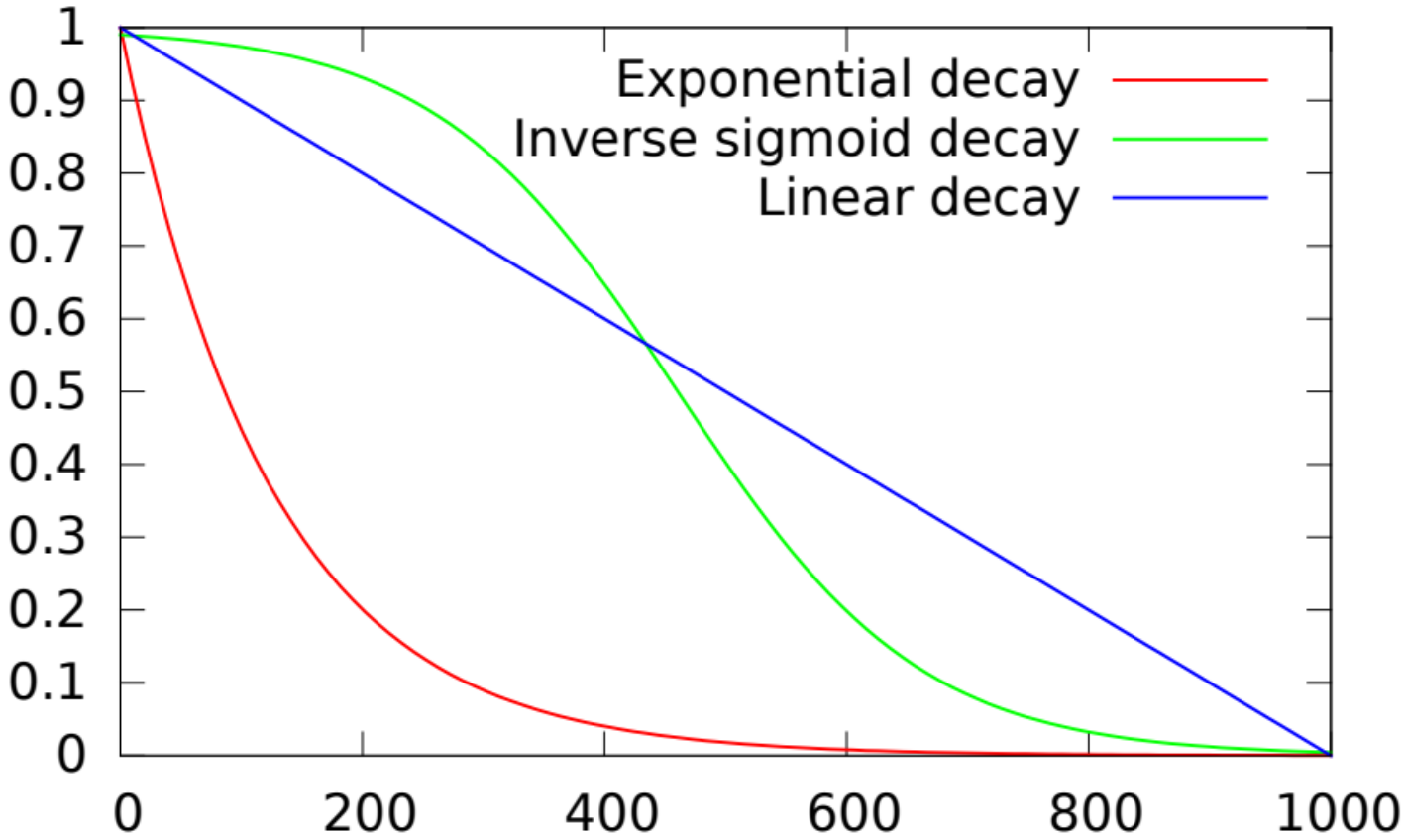
Механизм сэмплирования y_t i-го минибатча:

- С вероятностью ϵ_i используем y_{t-1}
- С вероятностью $1 - \epsilon_i$ используем $\hat{y}_{t-1}$

Уменьшаем ϵ_i от 1 до 0 по следующим расписаниям:

- Linear decay: $\epsilon_i = \max(\epsilon, k - ci)$
- Exponential decay: $\epsilon_i = k^i, i < 1$
- Inverse sigmoid decay: $\epsilon_i = \frac{k}{k + \exp(\frac{i}{k})}, k \geq 1$

Decay Schedules



Эксперименты

Image Captioning

Table 1: Various metrics (the higher the better) on the MSCOCO development set for the image captioning task.

Approach vs Metric	BLEU-4	METEOR	CIDER
Baseline	28.8	24.2	89.5
Baseline with Dropout	28.1	23.9	87.0
Always Sampling	11.2	15.7	49.7
Scheduled Sampling	**30.6**	**24.3**	**92.1**
Uniform Scheduled Sampling	29.2	24.2	90.9
Baseline ensemble of 10	30.7	25.1	95.7
Scheduled Sampling ensemble of 5	**32.3**	**25.4**	**98.7**

- Inverse sigmoid decay schedule for ϵ_i
- 2015 MSCOCO image captioning challenge: первое место

Эксперименты

Constituency Parsing

Table 2: F1 score (the higher the better) on the validation set of the parsing task.

Approach	F1
Baseline LSTM	86.54
Baseline LSTM with Dropout	87.0
Always Sampling	-
Scheduled Sampling	**88.08**
Scheduled Sampling with Dropout	**88.68**

- Inverse sigmoid decay schedule for ϵ_i

Эксперименты

Speech Recognition

Approach	ϵ_s	ϵ_e	Next Step FER	Decoding FER
Always Sampling	0	0	34.6	35.8
Scheduled Sampling 1	0.25	0	34.3	**34.5**
Scheduled Sampling 2	0.5	0	34.1	35.0
Scheduled Sampling 3	0.9	0.5	19.8	42.0
Baseline LSTM	1	1	15.0	46.0

- При тестировании используют beam search decoding (beam size 10)
- Linear decay schedule for ϵ_i

Идеи для будущей работы

- Способ обучения модели не точный - не учитываются градиенты вероятностей, с которыми сэмплируются токены
- Исследование лучших стратегий сэмплирования, в том числе используя уверенность модели

Выводы

- Стандартный способ обучения RNN отличается от того, как мы используем модель во время генерации, что приводит к накоплению ошибки на этапе тестирования
- Изменяется процедура обучения, во время которой каждый ground truth токен иногда заменяется на предыдущее предсказание модели
- Эксперименты показывают улучшение качества на этапе предсказания, не увеличивая время обучения

Статья

- Samy Bengio, Oriol Vinyals, Navdeep Jaitly, Noam Shazeer *Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks, 2015*