Fine-tuning neural conversation models for auxilary goals by means of deep reinforcement learning

Дмитрий Андреевич Персиянов

Московский физико-технический институт

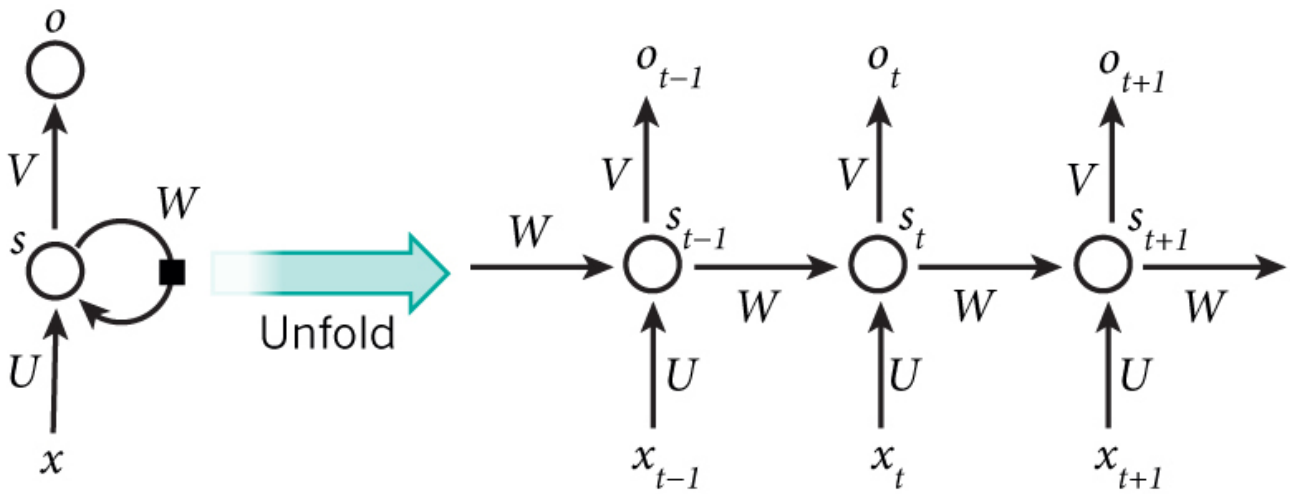
23 мая 2017 г.

План

- Conversational модели
- RL дообучение
- BePolite эксперимент
- BeLikeX эксперимент
- Заключение и дальнейшие исследования

Conversational модели

В последнее время рекуррентные сети успешно используется для построения языковых и sequence-to-sequence моделей. Обучение происходит на огромных корпусах текстов.



Имея обучающий пример $(\mathbf{c}, \mathbf{a})$ контекст-ответ, где $\mathbf{c} = \{c_1, c_2, \ldots, c_n\}$, $\mathbf{a} = \{a_1, a_2, \ldots, a_k\}$, учим модель, минимизируя лосс:

$$L(\theta) = -\sum_{t=1}^{k} \log \left(p_\theta(a_t | a_1, \ldots, a_{t-1}) \right)$$

или (в RL нотации)

$$L(\theta) = -\mathbb{E}_{\mathbf{a} \sim \mathcal{D}} \left[\log p_\theta(\mathbf{a}) \right]$$

Conversational модели, проблемы

- На один и тот же вопрос два разных ответа (inconsistency)
- Выучиваем, минимизируя кроссентропию, а нам иногда хочется другого:
  - Консистентность (учитывание контекста предыдущих ответов)
  - **Запрет на использование каких-то слов**
  - **Ведение беседы в каком-то стиле**
  - Максимизация скорости завершения диалога
  - Максимизация удовлетворенности пользователя
  - Максимизация ...

RL дообучение

Диалоговую модель $p_\theta(a_t|h_t, a_{t-1})$ можно воспринимать как политику $\pi_\theta(a_t|s_t)$.
Необходимо найти политику $\pi(a|s)$, такую что

$$\mathbb{E}_{\hat{\mathbf{a}}\sim\pi}\left[R_0 + \gamma R_1 + \cdots + \gamma^t R_t + \cdots\right] \to max,$$

где $R(\mathbf{a}, \hat{\mathbf{a}})$– некоторая функция награды, зависящая от правильного ответа $\mathbf{a}$ из обучающей выборки и сгенерированного моделью ответа $\hat{\mathbf{a}}$.
Также возможен более гранулярный вариант $R(a_t, \hat{a}_t)$.

BePolite эксперимент

- Данные: opensubtitles.org (en), 18млн пар (контекст, ответ).
- Собрали 800 обсценных слов (маты, религиозные/расовые оскорбления). Обозначим это множество за $\mathcal{S}$.
- Функция наград: $R(\hat{a}_t) = -\mathbb{I}[\hat{a}_t \in \mathcal{S}]$
- Используем предобученную по MLE лоссу модель.
- Дообучаем policy-gradient методом по $L(\theta) = -\mathbb{E}_{\hat{\mathbf{a}} \sim p_\theta}\left[\sum_{t=1}^{k} R(\hat{a}_t) \log p_\theta(\hat{a}_t | \hat{a}_{t-1}, \dots)\right] - \alpha \mathbb{E}_{\mathbf{a} \sim \mathcal{D}}\left[\log p_\theta(\mathbf{a})\right]$
- $\alpha = 5, 20$.
- Обучаем 500 батчей по 64 примера (около 30 минут).

BePolite эксперимент

Таблица: Метрики бейзлайна

Средняя награда	Перплексия
-0.136	3.142

Таблица: Метрики после policy-gradient дообучения

α	Средняя награда	Перплексия
5	-0.021	3.297
20	-0.065	3.270

BeLikeX эксперимент

- Данные: twitter (ru), 50млн примеров (контекст, ответ) + каждое сообщение размечено id пользователя.
- Отобрали 1000 пользователей по частоте участия в диалогах. (Топ1 – 9500 ответов на чьи-то твиты).
- Обучили dssm-like модель $D(\mathbf{uid}, \mathbf{a}) \in [-1, 1]$ в качестве прокси-награды.
- Выбрали одного юзера с большим кол-вом сообщений.

BeLikeX эксперимент

Таблица: Метрики на валидационных выборках

Модель	Перплексия	Перплексия/uid	Средняя награда
baseline	4.235	5.249	0.258
llh on user	5.792	6.540	0.389
dssm weighting	4.337	5.358	0.281
RL-finetuned	?	?	?

Похожие работы

- Deep Reinforcement Learning for Dialogue Generation (https://arxiv.org/pdf/1612.00563.pdf) – дообучают RL-ом, но борются с проблемой затухания диалогов и общих ответов.
- A Persona-Based Neural Conversation Model (https://nlp.stanford.edu/pubs/jiwei2016Persona.pdf) – выучивают ембеддинги для пользователей и подают на вход декодеру.

Заключение и дальнейшие исследования

- RL помогает быстро и эффективно дообучать модели под разные требования, выразимые в виде функции наград.

- BePolite: посмотреть как запрет одних слов влияет на частоту использования семантически близких, но которых нет в словаре
- BeLikeX: использовать дискриминатор, обученный лишь на одном юзере, как в GAN'ах. Пытаться обмануть его.