Imitation Learning in Dialogue Generation

Wentao Qin

May 17, 2019

Content

- Three Approach to Imitation Learning
  - Behavior Cloning
  - Inverse Reinforcement Learning
  - Generative Adversarial Imitation Learning
- Imitaion Learning in Dialogue Generation

Characteristic of Imitation Learning

- Data Driven, learn to perform a task from expert demonstrations
    - Imitate what expert(human) does: Given State s, take right action a

Context(s)：
Speak 1: Long time no see
Speak 2: Yup, how are you?

Response(a)：
Answer 1: I am good!
Answer 2: I don’t know.
Answer 3: no, no, no

Behavior Cloning

- Supervised Style , one way to achieve Imitation Learning
- Approach: Model that can learn a mapping from state to a good action, by optimizing a well designed object function.
  - All Supervised Method can be seen as behavior cloning
- In dialogue generation, Ses2Seq model is a popular method. By applying negtive log likelihood loss function in each decoding step, make the machine to clone/imitate human responses

The Disadvantages of Behavior Cloning

- The set of Observed states is limited. Therefore, when faced with a (totally) new state, the taken action may cause failure
  - Solution: Data aggregation, i.e., providing more traning data
- Behavior cloning treats every sample in trainging data equally, which cause problem of generic response in dialogue generation
  - Soulution: Design new architecture and new object function. Reinforcement Learning comes up.

Reinforcement Learning in Dialogue generation

[1] Deep Reinforcement Learning for Dialogue Generation, Li et al, 2016

Common loss function (log likelihood based):

$$loss = -\sum log(a|s)$$

RL loss function (with reward as weight, policy gradient style)

$$RLloss = -\sum r(s,a) log(a|s)$$

$r(s,a)$ is reward when taking action a under state s

In [1], the authors manually desgin reward function based on 3 aspects:

1. Ease of answering
2. New information
3. Semantic Coherence

Imitation Learning using Inverse Reinforcement Learning

Reinforcement Learning has a severe problem —— It's hard to design a good reward function.

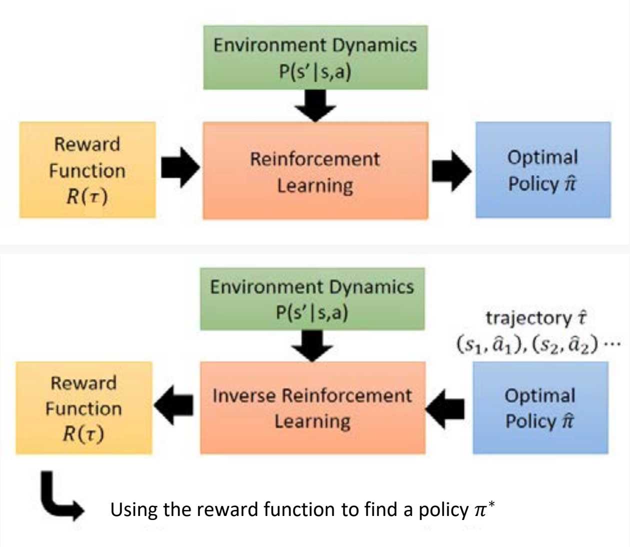
- For auto pilot, how much the reward of running through the red light, of hitting other cars. Hard to decide.
- For dialogue generation, there is no a gloden metrics to judge whether a response is good or not. The existing researchs manually design reward in certain aspects. The designed reward may let the machine work well in these ascpects, but not a general solution.

Imitation Learning using Inverse Reinforcement Learning

Inverse Reinforcemnt Learning: a second kind of imitation learning

- Based on Reinforcemnt Learning
- Key idea: **learn the reward function**

Reinforcement Learning v.s. Inverse Reinforcement Learning



Inverse Reinforcement Learning

Question：How to get the reward function?

Solutions:

1. Maximum entropy model
2. Structure Perceptron

Key Assumption: for $(s, \bar{a}) \in \pi_E(Training\, data)$, $(s, a) \in \pi(Sampled\, form\, cueent\, policy)$, the reward function $r(s, a)$ should satisfy $r(s, a) \leq r(s, \bar{a})$

Maximun Entropy Inverse Reinforcement Learning

Update reward fuction by minimize:

$$L = -\mathbb{E}_{(s,\bar{a})\in\pi_E}[r(s,\bar{a})] + \log(\mathbb{E}_{(s,a)\in\pi}\left[\frac{e^{r(s,a)}}{\pi(s,a)}\right])$$

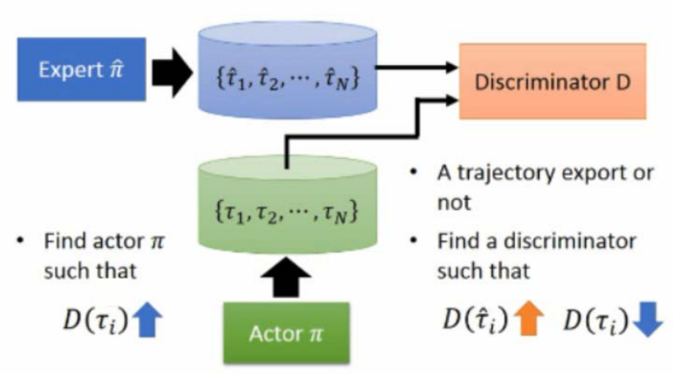
For more details, see *Maximum entropy inversereinforcement learning. AAAI 2008*

Inverse Reinforcement Learning

- Updating the reward function and runing the standard reinforcment learning are taken alternately, util convergence
- Disadvantage: It is extremely expensive to run
    - This procedure requires reinforcement learning in every inner loop, which is slow

Generative Adversarial Imitation Learning: GAN-style

- Key idea: using the Discriminator to signal reward



Imitation Learning in Dialogue Generation

Adversarial Learning for Neural Dialogue Generation

ACL 2017

Jiwei Li, Will Monroe, Tianlin Shi, Sebastien Jean, Alan Ritter and Dan Jurafsky

Dialogue Generation: From Imitation Learning to Inverse Reinforcement Learning

AAAI 2019

Ziming Li, Julia Kiseleva, and Maarten de Rijke

Imitation Learning in Dialogue Generation

Problem Setting

- Response $< w_1, w_2, \dots, w_t >$ can be regarded as corresponding actions $< a_1, a_2, \dots, a_n >$ at different steps
- Use a state function f to compress the dialogue context p and the words already generated
  - $s_1 = f(p)$
  - $s_t = f(p, a_1, a_2, \dots, a_{t-1})$
- Find optimal policy $\pi(a_t|s_t)$ that selects the most appropriate word at each time step

Adversarial Imitation Learning

1. Discriminator：
    - A hierachical structure to compress utterances

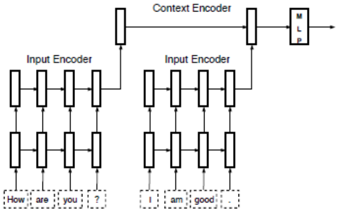


Figure 1: Discriminator architecture in DG-AIL.

    - Minimize $\mathbb{E}_{\pi}\left[log\left(D(s,a)\right)\right] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))]$

What SeqGANs do!

Adversarial Imitation Learning

2. Generator:
   - A Seq2Seq model
   - At each decode step t, choose an action a_t(i.e., chooes a word w_t), then use the Discriminator to compute $D(s_t, a_t)$ as reward, then use policy gradient to update Generator's parameters
   - Maximize $\mathbb{E}_\pi\left[\sum_t log\left(D(s_t, a_t)\right)\right]$
   - Note: The discriminator is trained to assign scores for fully generated sequences, while the action a_t in intermediate steps only represents partially decoded sequences
     - Solution 1: Discriminator assigns rewards to both fully and partially decoded sequences
     - Solution 2: Use Monte Carlo search to get several full sequences, and the average score as reward

Regularation Trick ---- Maximum causal entropy

The causal entropy of policy π

$$H(\pi) = \mathbb{E}_{\pi}[-\,log\,\pi(a|s)]$$

- It measures the uncertainy presented in π
- In learning a probability model, among all possible models, model with max entropy is the best one.
  - Possible model: satisfy existing data
  - Max entropy: don't make any subjective assumptions about the unseen data
- New Objective function of Generator:
- Maximize $\lambda H(\pi) + \mathbb{E}_{\pi}\left[\sum_t log\left(D(s_t, a_t)\right)\right]$

Imitation Learning in Dialogue Generation

Maximum entropy Inverse reinforcement learning

1. Reward function model：
    - A hierachical structure to compress state and action, and then a MLP layer is used to get a scalar as reward

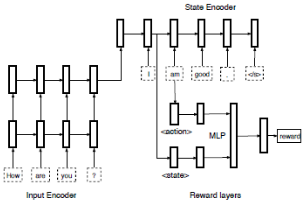


Figure 2: Reward model architecture in DG-AIRL.

Minimize $-\mathbb{E}_{(s,\bar{a})\in\pi_E}[r(s,\bar{a})] + \log(\mathbb{E}_{(s,a)\in\pi}\left[\frac{e^{r(s,a)}}{\pi(s,a)}\right])$

Imitation Learning in Dialogue Generation

Maximum entropy Inverse reinforcement learning

2. Dialogue response policy：
   - Same with previous practice
   - Difference lies in use the reward function model to get reward
   - Maximize $\lambda H(\pi) + \mathbb{E}_{\pi}\left[\sum_t log\left(r(s_t, a_t)\right)\right]$

Experiment

- Dataset: MovieTripe dataset
    - Train:valid:test = 157000:19000:19000
    - Vocablary size = 20k
    - Embedding size = 200
- Baseline
    - Seq2Seq+Attention
    - VHERD
    - SeqGan
    - DG-AIL
        - SeqGAN with maximum causal entropy
    - DG-AIRL

Experiment

- Existing Automatic Evaluation metrics
  - BLEU
  - Embedding metrics
    - Average embedding
    - Greedy embedding
    - Extrema embedding
  - Distinct

- Chosen evaluation metric: Embedding metrics
  - Why not BLEU: Word-overlap metrics such as BLEU correlate very weekly with reply quality judgements from human annotators
  - Why not Distinct: The authors found that the Distinct result is not aligned with the results besed on human evaluations

Experiment

- Embedding metrics

Model	Average	Greedy	Extrema	Length
Seq2Seq	0.563 ± 0.003	0.167 ± 0.001	0.352 ± 0.002	8.8
SeqGan	0.564 ± 0.003	0.165 ± 0.001	0.354 ± 0.002	9.7
VHRED	0.507 ± 0.003	0.145 ± 0.001	0.309 ± 0.002	**12.0**
DG-AIL	0.553 ± 0.003	$\mathbf{0.171^* \pm 0.001}$	0.356 ± 0.002	7.7
DG-AIRL	$\mathbf{0.589^* \pm 0.003}$	0.169 ± 0.001	$\mathbf{0.368^* \pm 0.002}$	10

Table 1: Performance in terms of embedding metrics of response generation models, with 95% confidence intervals. * indicates the result is statistically significant ($p < 0.005$) with a paired t-test over DG-AIRL and other baseline models.

Experiment

- Human evaluations
  - Pairwise evaluation: given two models' result, ask human which is better based on the following aspects, tie is allowed
    - (Top priority) is relevent?
    - Is natural?
    - Is interesting?
    - Is proactive, i.e., can make conversation continue?
    - Is the only possible reply to the given context?

Model pair	Win	Tie	Loss
DG-AIRL-Seq2Seq	**0.44**	0.29	0.27
DG-AIRL-VHRED	**0.46**	0.32	0.22
DG-AIRL-SeqGan	**0.47**	0.25	0.28
DG-AIRL-DG-AIL	0.36	**0.37**	0.27

Table 2: Performance in terms of pairwise human annotations of response generation models.

Experiment

- Human evaluations
  - Pointwise evaluation: ask human to score response among 0, +1, +2
    - +2 (a) relevent, natrual, informative, interesting;
      (b) natural, make the conversation continue
      (c) the only possible reply to the context
    - +1 can be used as a reply to the context, but is too generic like “I don’t know”, which usually is reactive
    - 0 cannot be a reply to the context. Either semantically irrelevant or disfluent

Experiment

- Ponitwise evaluation

Model	Freq of +2	Freq of +1	Freq of 0	Avg Score
Seq2Seq	0.09	0.22	0.69	0.40
SeqGan	0.09	0.21	0.70	0.39
VHRED	0.12	0.25	0.63	0.49
DG-AIL	0.12	0.29	0.59	0.53
DG-AIRL	0.13	0.28	0.59	**0.54**

Table 3: Performance in terms of pointwise human evaluations of response generation models. "Freq of N" is the relative frequency of a model's responses with a score of N.

Some thinking and Q&A

- Why reinforcemnt learning?
- Why Imitation learning, like inverse reinforcement learning, for adversarial learning?
- What these learning methods actually do?

"The seemingly simple reward function can guide a complex (maybe powerful) strategy."

Reference & Extra reading

[1]Deep Reinforcement Learning for Dialogue Generation, Li et al, 2016

[2] Dialogue Generation: From Imitation Learning to Inverse Reinforcement Learning, AAAI, 2019 (IM+DG)

[3] Adversarial Learning for Neural Dialogue Generation (SeqGan+DG)

[4] Generative Adversarial Imitation Learning, OpenAI, 2016

[5] Imitation Learning with Recurrent Neural Networkss

Thanks for listening!