Adversarial Learning for Neural Dialogue Generation

Zvengin
zvengin@nii.ac.jp

December 30, 2017

1 General Introduction

1.1 Problem

As response generated by conventional sequence-to-sequence model which use MLE optimization method is dull, repetitive, generic and short-sighted, the author hopes to generate an indistinguishable response from human beings.

1.2 Methods

The dull,repetitive, generic and short-sighted response is caused by simplified MLE objective function which we use to optimize model's parameters in conventional sequence-to-sequence model. Therefore, author adopts reward instead of objective function, in order to avoid the influence of choosing an inappropriate objective function. Hence, reinforcement learning combined with adversarial training are used to improve the quality of response. The entire model is comprised of two main parts, one is generator and the other is discriminator. Then reinforcement learning method is applied to the entire model with the predication of discriminator as reward. Actually generator in essence is a sequence-to-sequence model and discriminator is a binary classifier. At the begining, dialogue history is inputted into generator and generator will generates a response. Then generated response combined with dailogue history is fedded into discriminator Finally, the prediction of discriminator will feed back to generator for updating model's parameters.

2 Details

1. Author adopts sequence-to-sequence model as response generator
2. Before inputting dailogue history and response into discriminator, a hierarchical encoder is used to encode dialogue history and response into a vector.
3. Policy gradient methods is applied to training this model. The prediction, that response is generated by human, of discriminator is feeded back to generator as reward. The training goal is to maxmize the expected reward of generated response, $J(\theta) = E_{y\sim p(y|x)}(Q_+(\{x,y\})|\theta)$, using reinforcement algorithm. Author also use liklihood ratio trick to approximate gradient.

$$\begin{aligned} \nabla J(\theta) &\approx [Q_+(\{x,y\}) - b(\{x,y\})]\nabla log\pi(y|x) && (1) \\ &= [Q_+(\{x,y\}) - b(\{x,y\})]\sum_t \nabla logP(y_t|x, y_{1:t-1}) && (2) \end{aligned}$$

4. As we can see, in equation (2) we find that vanilla Reinforce Model assigns the same negative reward to all tokens within human-generated response, whereas proper credit assignment in training would give seperate rewards. author call this reward for every generation step, abbreviated REGS.
5. As discriminator can't score a partial response, author propose two solutions to this problem. One is Monte Caarlo, given a partial response Sp, we continue sampling from response's distribution untill we get complete response and repeat this process for N times. Finally we get N responses. The average score of N responses is used as reward for partial response Sp. The drawback of this method is time consuming. The other convert a fully response discriminator to a partial response discriminator. By adopting REGS, the approximated gradient is

$$\nabla J(\theta) \approx \sum_{t} (Q_{+}(x, y_t) - b(x, y_t)) \nabla log P(y_t | x, y_{1:t-1}) \tag{3}$$