Adversarial Learning for Neural Dialogue Generation —— hoho

论文试图解决什么问题？

现有大多数对话生成方法对训练的目标过于简单化（如只用最大似然估计），导致往往很容易生成一些无意义的、重复的、无聊的语言表达。

这是否是一个新的问题？

不是

这篇文章要验证一个什么科学假设？

是否可以仿照图灵测试的观点，使用一个判别方法判定机器生成的对话跟人类生成的对话是否相像。是则认为这个生成对话是OK的

有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员？

to do

论文中提到的解决方案之关键是什么？

核心：参考GAN的方法，使用generator生成对话，使用driminator作为判别器，衡量生成的对话的reward，然后用强化学习的方式，将生成对话的过程建模为MDP，使用PG方法进行生成策略的提升。

基本流程：

1. Generator: 使用Seq2Seq模型，将输入的历史语句x编码为向量，然后用softmax计算每个token生成的概率，action为生成每个token，最终形成输出序列y （这样称为一个episode）
2. Discriminator：这是二分类模型，将{x, y}作为输入，使用分层的encoder将其编码，使用二分类softmax输出是否是人类生成语言的概率$Q_+(\{x,y\})$或机器生成语言的概率$Q_-(\{x,y\})$
3. 将$Q_+(\{x,y\})$作为Generator的reward，使用REINFORCEMENT算法，最大化Generator的期望回报$J(\theta) = \mathbb{E}_{y\sim p(y|x)}(Q_+(\{(x,y)\})|\theta)$

每个episode生成的序列y中的每个token（即动作的序列）都赋予同一个reward，为了解决这个credit assignment问题（譬如ground truth是”I am John”，但生成的序列为“I don’t know”，合理reward分配方式为“I”赋予正reward，“don’t”和“know”赋予负的reward，而不是3个token都统一赋予相同的reward），作者采用两种方案：

（1）Monte Carlo方法

给定前缀s_p，模型持续生成后续token直到结束，重复这个过程N遍，得到N个采样序列，然后这N个采样序列输入到discriminator，将输入在reward作平均，作为s_p的最终reward。显然这种方法训练效率较低。

（2）训练discriminator使其能够单独给每个token赋予特定的reward

将生成序列分为正样本集合$\{y^+_{1:t}\}$和负样本集合$\{y^-_{1:t}\}$（为ground truth就是正样本，否则归为负样本），分别从两个集合中随机采样一个样本训练discriminator，这种方法虽然效率高但准确度低。

最后作者还是采取MC方法，PG更新方式为：

$$\nabla J(\theta) \approx \sum_t (Q_+(x, Y_t) - b(x, Y_t)) \nabla \log p(y_t|x, Y_{1:t-1})$$

$Y_t = y_{1:t}$

$b(x, Y_t)$为基本的REINFORCE模型

为了进一步提高模型性能，还采取了teacher forceing的方式，直接将人类生成的序列输入到generator，此时discriminator输出的reward设置为1（或其他正值），然generator进行更新

整个算法流程如下：

For number of training iterations **do**
. **For** i=1,D-steps **do**
. Sample (X,Y) from real data
. Sample $\hat{Y} \sim G(\cdot|X)$
. Update D using (X, Y) as positive examples and $(X, \hat{Y})$ as negative examples.
. **End**
.
. **For** i=1,G-steps **do**
. Sample (X,Y) from real data
. Sample $\hat{Y} \sim G(\cdot|X)$
. Compute Reward r for $(X, \hat{Y})$ using D.
. Update G on $(X, \hat{Y})$ using reward r
. Teacher-Forcing: Update G on (X, Y)
. **End**
End

论文中的实验是如何设计的？

to do

用于定量评估的数据集是什么？代码有没有开源？

無

论文中的实验及结果有没有很好地支持需要验证的科学假设？

to do

这篇论文到底有什么贡献？

“作者在最后的总结中写道这种强化学习的方法可以应用在很多NLP的生成任务中，但是在一些领域比如机器翻译，作者并没有得到明显的效果提升，原因可能在于这种方法更适用于target的熵更高的任务中。仔细想想这也是可以理解的，只有target更丰富生成器的探索才会更有意义，reward蕴含的信息也越多。如果target较为固定，比如标准输出只有一种形式，那么也没必要用reward了，直接用监督学习就可以得到较好的结果。”

下一步呢？有什么工作可以继续深入？

to do