ADL—Other RL

2019年4月23日　　下午 12:13

RL可以看成經歷一連串決策最後得到的reward跟答案
可以規避無法微分的情況
EX: sequence GAN

NLG
當training跟testing的metric會有差距
BLEU score不可微分
產生一段句子就是sequence of words，一連串的action所得到的回饋
把objective變成blue score，把loss定義乘negative reward
較一個人來評估好不好，zpgj41j4dk3j分數不可微分但是期望值分數可以微分，變成positive gradient

把cross-entropy跟RL合起來，MIXER paper:
先用cross-entropy去train不要太爛，再去用BLEU score
根據一整個句子去做learning，word level去train會有限制

dialogue generation MLE-based SEQ2SEQ:

- 希望可以有意義對話
- 利用Mutual Information來train
- 看著上下文，給定一句話，當給定這句話又可以infer回上下文，代表資訊量足夠
- Reward 可以自己設計，當應用到想要用的情境就可以自己設計
- 當產生這句話用I dont know可以回答就是不好
- 希望資訊量不要重疊(information flow，hidden state的cosine similarity要低)
- MI

Dual Learning for Machine Translation

- 給定task X給Y，找到另外一個task Y給X
- EX: ASR(語音辨識) vs TTS(text-to-speech)、翻譯
- 利用leverage unlabel data
- 資料1轉換成資料2，再把資料2轉換成資料1，是否一樣。中間產物是否夠好
    - 先supervised train一下比較好
- 中間產物: LM可以知道是否通順

理解、決策、生成 EndtoEnd Task-Completion Neural Dialouge system

- 從chat bot得到資訊，希望可以從很短term得到正確答案
- RL 不stable又難train

用RL去train，dialouge system不切實際，所以會變成收集一大堆聲音檔，if else去寫rule
但仍然存在gap，行為可能差勁，因為simulator無法cover真實行為
user用learn的，用NN去學world model，當成環境
Deep Dyna-Q，去學習environment。因為跟真實環境互動很貴，所以假環境去互動
因為禁不起一次失敗
但如果NN不好，可能會損害policy，假data不夠好，所以應該要有classifier
=> Discriminative Deep Dyna-Q，只留下高品質的假data