BLEU

강사 : 백병인

pi.paek@modulabs.co.kr

모두의연구소 Research Scientist

2019 모두의연구소

Decoder Evaluation Metric

- Decoder 의 학습(train)시 loss는 어떻게 평가할까?
  - Image Decoder 라면 => Target Image와의 Mean Square Error 최소화
  - Text Decoder라면 => Target Text에서의 next word와 decoder가 구현하고 있는 language model 사이의 cross entropy 최소화

- Decoder 생성 결과의 품질을 어떻게 측정할 것인가?
  - test 단계엔 target이 없다. 무엇과 비교하는 방식을 적용할 수 없다.
  - MOS (Mean Opinion Score) - 정성적 방법
- 정량적인 방법은?
  - PSNR(Peak Signal to Noise Ratio), SSIM(Structural SIMilarity)
  - BLEU(BiLingual Evaluation Understudy)

BLEU(BiLingual Evaluation Understudy)

- Machine Translation 결과물의 품질 측정에 널리 사용되는 metric
  - n-gram을 통한 순서쌍들이 얼마나 겹치는지 측정(precision)
  - 문장길이에 대한 과적합 보정 (Brevity Penalty)
  - 같은 단어가 연속적으로 나올때 과적합 되는 것을 보정(Clipping)

$$BLEU = min(1, \frac{output\ length(\text{예측 문장})}{reference\ length(\text{실제 문장})})(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}}$$

https://donghwa-kim.github.io/BLEU.html

Precision

- n-gram(1~4)을 통한 순서쌍들이 얼마나 겹치는지 측정

  - `예측된 sentence` : **빛이 쐬는** 노인은 **완벽한** 어두운곳에서 **잠든 사람과 비교할 때** 강박증이 **심해질** 기회가 **훨씬 높았다**
  - `true sentence` : **빛이 쐬는** 사람은 **완벽한** 어둠에서 **잠든 사람과 비교할 때** 우울증이 **심해질** 가능성이 **훨씬 높았다**

- 1-gram precision: $\frac{\text{일치하는 } 1\text{-}gram\text{의 수 (예측된 } sentence\text{중에서)}}{\text{모든 } 1\text{-}gram\text{쌍 (예측된 } sentence\text{중에서)}} = \frac{10}{14}$
- 2-gram precision: $\frac{\text{일치하는 } 2\text{-}gram\text{의 수 (예측된 } sentence\text{중에서)}}{\text{모든 } 2\text{-}gram\text{쌍 (예측된 } sentence\text{중에서)}} = \frac{5}{13}$
- 3-gram precision: $\frac{\text{일치하는 } 3\text{-}gram\text{의 수 (예측된 } sentence\text{중에서)}}{\text{모든 } 3\text{-}gram\text{쌍 (예측된 } sentence\text{중에서)}} = \frac{2}{12}$
- 4-gram precision: $\frac{\text{일치하는 } 4\text{-}gram\text{의 수 (예측된 } sentence\text{중에서)}}{\text{모든 } 4\text{-}gram\text{쌍 (예측된 } sentence\text{중에서)}} = \frac{1}{11}$

$$(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}} = (\frac{10}{14} \times \frac{5}{13} \times \frac{2}{12} \times \frac{1}{11})^{\frac{1}{4}}$$

https://donghwa-kim.github.io/BLEU.html

Brevity Penalty

- 문장길이에 대한 과적합 보정

  - `예측된 sentence` : **빛이 쐬는** 노인은 **완벽한** 어두운곳에서 잠듬
  - `true sentence` : **빛이 쐬는** 사람은 **완벽한** 어둠에서 잠든 사람과 비교할 때 우울증이 심해질 가능성이 훨씬 높았다

$$min(1, \frac{\text{예측된 } sentence\text{의 길이(단어의 갯수)}}{true\ sentence\text{의 길이(단어의 갯수)}}) = min(1, \frac{6}{14}) = \frac{3}{7}$$

https://donghwa-kim.github.io/BLEU.html

Clipping

- 같은 단어가 연속적으로 나올때 과적합 되는 것을 보정

  - `예측된 sentence` : The more decomposition the more flavor the food has
  - `true sentence` : The more the merrier I always say

  - 1-gram precision: $\frac{\text{일치하는} 1-gram\text{의 수(예측된 } sentence\text{중에서)}}{\text{모든} 1-gram\text{쌍 (예측된 } sentence\text{중에서)}} = \frac{5}{9}$
  - (보정 후) 1-gram precision: $\frac{\text{일치하는} 1-gram\text{의 수(예측된 } sentence\text{중에서)}}{\text{모든} 1-gram\text{쌍 (예측된 } sentence\text{중에서)}} = \frac{3}{9}$

https://donghwa-kim.github.io/BLEU.html

BLEU is true metric?

- 생성된 sentence : 나는 어제 집에 가서 잠을 잤다.
- True sentence : 나는 어제 집에 가서 잠을 설쳤다.

- 엄밀하게, BLEU는 번역 품질을 나타내는 true metric은 아니다.
- 그러나 BLEU가 높을 수록 NMT 모델의 성능이 좋을 가능성이 높으므로 통계적인 지표로 삼을 수는 있다.

2019 모두의연구소