Minimum Risk Training

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Motivations

Cannot reflect generation quality

- PPL은 정확한 생성 품질을 알 수 없음

- PG는 보상 함수의 미분이 필요 없어, BLEU를 통해 최적화 할 수 있음

Remove Teacher Forcing

- NLG는 auto-regressive task이므로 teacher forcing을 통해 학습함
  - 학습과 추론 사이의 괴리가 발생

- RL은 샘플링 기반의 학습이므로, 학습과 추론 방법의 차이가 없음

Equations

- Define risk:

$$\mathcal{D} = \{x^i, y^i\}_{i=1}^N$$

$$\begin{aligned}\mathcal{R}(\theta) &= \sum_{i=1}^N \mathbb{E}_{\hat{y}\sim P(\text{y}|x^i;\theta)}[\Delta(\hat{y}, y^i)] \\ &= \sum_{i=1}^N \sum_{\hat{y}\in\mathcal{Y}(x^i)} P(\hat{y}|x^i;\theta)\Delta(\hat{y}, y^i)\end{aligned}$$

$$\hat{\theta}_{\text{MRT}} = \underset{\theta\in\Theta}{\text{argmin}}\, \mathcal{R}(\theta)$$

Equations

- Re-define risk:

$$\tilde{\mathcal{R}}(\theta) = \sum_{i=1}^{N} \mathbb{E}_{\hat{y} \sim Q(\mathrm{y}|x^i;\theta,\alpha)}[\Delta(\hat{y}, y^i)]$$

$$= \sum_{i=1}^{N} \sum_{\hat{y} \in \mathcal{S}(x^i)} Q(\hat{y}|x^i;\theta,\alpha)\Delta(\hat{y}, y^i)$$

where $\mathcal{S}(x^i)$ is a sampled subset of the full search space $\mathcal{Y}(x^i)$,
and $Q(\hat{y}|x^i;\theta,\alpha)$ is a distribution defined on the subspace $\mathcal{S}(x^i)$:

$$Q(\hat{y}|x^i;\theta,\alpha) = \frac{P(\hat{y}|x^i;\theta)^\alpha}{\sum_{y' \in \mathcal{S}(x^i)} P(y'|x^i;\theta)^\alpha}.$$

Equations

- Calculate gradients:

$$\begin{aligned}\nabla_\theta\tilde{\mathcal{R}}(\theta) &= \alpha\sum_{i=1}^{N}\mathbb{E}_{\hat{y}\sim P(\mathrm{y}|x^i;\theta)^\alpha}\Big[\frac{\nabla_\theta P(\hat{y}|x^i;\theta)}{P(\hat{y}|x^i;\theta)}\times\big(\Delta(\hat{y},y^i)-\mathbb{E}_{y'\sim P(\mathrm{y}|x^i;\theta)^\alpha}[\Delta(y',y^i)]\big)\Big] \\ &= \alpha\sum_{i=1}^{N}\mathbb{E}_{\hat{y}\sim P(\mathrm{y}|x^i;\theta)^\alpha}\Big[\nabla_\theta\log P(\hat{y}|x^i;\theta)\times\big(\Delta(\hat{y},y^i)-\mathbb{E}_{y'\sim P(\mathrm{y}|x^i;\theta)^\alpha}[\Delta(y',y^i)]\big)\Big] \\ &\approx \alpha\sum_{i=1}^{N}\nabla_\theta\log P(\hat{y}|x^i;\theta)\times\big(\Delta(\hat{y},y^i)-\frac{1}{K}\sum_{k=1}^{K}\Delta(y^k,y^i)\big),\text{ where }\hat{y}\sim P(\mathrm{y}|x^i;\theta)^\alpha.\end{aligned}$$

$$\theta\leftarrow\theta-\eta\nabla_\theta\tilde{\mathcal{R}}(\theta)$$

Evaluations

System	Architecture	Training	Vocab	BLEU
Existing end-to-end NMT systems				
Bahdanau et al. (2015)	gated RNN with search	MLE	30K	28.45
Jean et al. (2015)	gated RNN with search		30K	29.97
Jean et al. (2015)	gated RNN with search + PosUnk		30K	33.08
Luong et al. (2015b)	LSTM with 4 layers		40K	29.50
Luong et al. (2015b)	LSTM with 4 layers + PosUnk		40K	31.80
Luong et al. (2015b)	LSTM with 6 layers		40K	30.40
Luong et al. (2015b)	LSTM with 6 layers + PosUnk		40K	32.70
Sutskever et al. (2014)	LSTM with 4 layers		80K	30.59
Our end-to-end NMT systems				
this work	gated RNN with search	MLE	30K	29.88
	gated RNN with search	MRT	30K	31.30
	gated RNN with search + PosUnk	MRT	30K	34.23

Table 7: Comparison with previous work on English-French translation. The BLEU scores are case-sensitive. "PosUnk" denotes Luong et al. (2015b)'s technique of handling rare words.

[Shen et al., 2015]

RL in GNMT

- Minimize both MLE and RL loss:

$$\mathcal{O}_{ML}(\theta) = \sum_{i=1}^{N} \log P_\theta(Y^{*(i)}|X^{(i)})$$

$$\mathcal{O}_{RL}(\theta) = \sum_{i=1}^{N} \sum_{Y \in \mathcal{Y}} P_\theta(Y|X^{(i)}) r(Y, Y^{*(i)})$$

$$\mathcal{O}_{Mixed}(\theta) = \alpha * \mathcal{O}_{ML}(\theta) + \mathcal{O}_{RL}(\theta)$$

Table 6: Single model test BLEU scores, averaged over 8 runs, on WMT En→Fr and En→De

Dataset	Trained with log-likelihood	Refined with RL
En→Fr	38.95	39.92
En→De	24.67	24.60

Summary

- MRT는 reward 대신 risk를 정의하여, minimize 문제로 만듦
  - 수식을 풀어 부호를 없애면, 결국 같은 수식
- Reward(or risk) 함수를 설계할 때, 매우 신중해야 함
  - 모델은 아주 작은 빈 틈도 파고들어, cheating을 시도할 것
  - GNMT는 BLEU를 개선한 GLEU를 만들어, 적용했다고 밝힘
- Monte-Carlo에 의해서 기대값은 샘플링의 평균값으로 대체
  - 심지어 샘플링 횟수가 1회이더라도 동작할 것
  - 샘플링에 의존하므로, 효율이 저하되는 것은 또 다른 문제