A Brief Introduction to Policy Gradient Method

朱小天

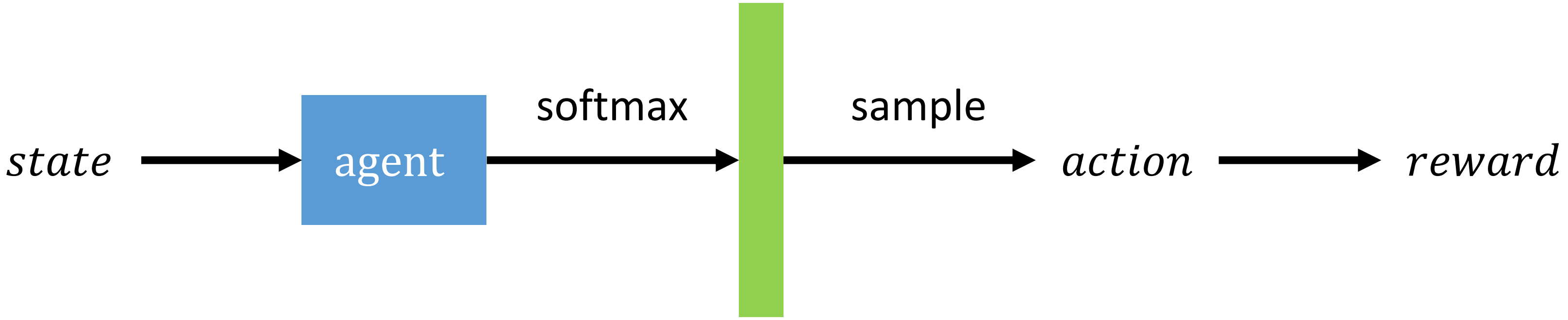
2018.10.12

Outline

- A brief introduction to policy gradient theorem
- Example: Learning Globally Optimized Object Detector via Policy Gradient
- Example: Towards Diverse and Natural Image Descriptions via a Conditional GAN

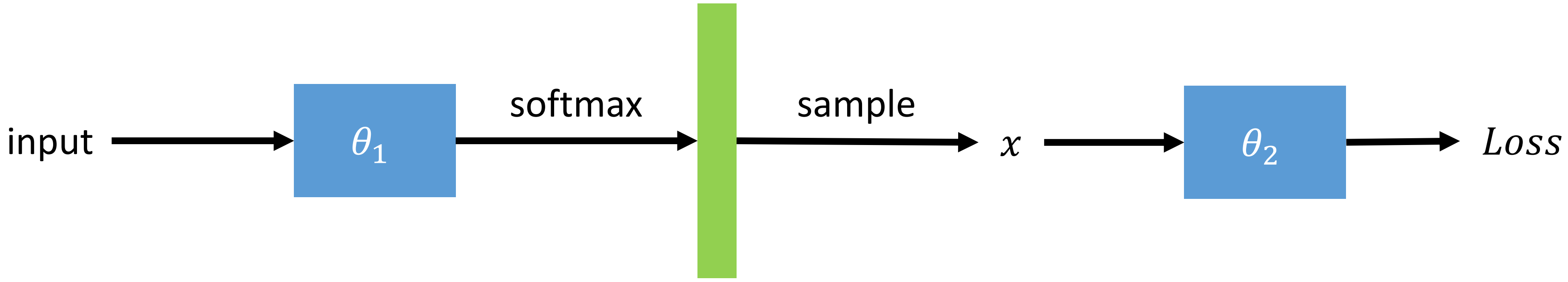
A brief introduction to policy gradient theorem

- 如何根据reward更新网络参数？



A brief introduction to policy gradient theorem

- 如何对采样求导？



$$\frac{\partial Loss}{\partial \theta_1} = ?$$

A brief introduction to policy gradient theorem

- We might be given a parameterized probability distribution $x \sim p(\cdot;\ \theta)$. In this case, we can use the *score function* (SF) estimator [3]:

$$\frac{\partial}{\partial \theta}\mathbb{E}_x\left[f(x)\right] = \mathbb{E}_x\left[f(x)\frac{\partial}{\partial \theta}\log p(x;\ \theta)\right]. \tag{1}$$

This classic equation is derived as follows:

$$\begin{aligned}\frac{\partial}{\partial \theta}\mathbb{E}_x\left[f(x)\right] &= \frac{\partial}{\partial \theta}\int dx\ p(x;\ \theta)f(x) = \int dx\ \frac{\partial}{\partial \theta}p(x;\ \theta)f(x) \\ &= \int dx\ p(x;\ \theta)\frac{\partial}{\partial \theta}\log p(x;\ \theta)f(x) = \mathbb{E}_x\left[f(x)\frac{\partial}{\partial \theta}\log p(x;\ \theta)\right]. \end{aligned} \tag{2}$$

This equation is valid if and only if $p(x;\ \theta)$ is a continuous function of θ; however, it does not need to be a continuous function of x [4].

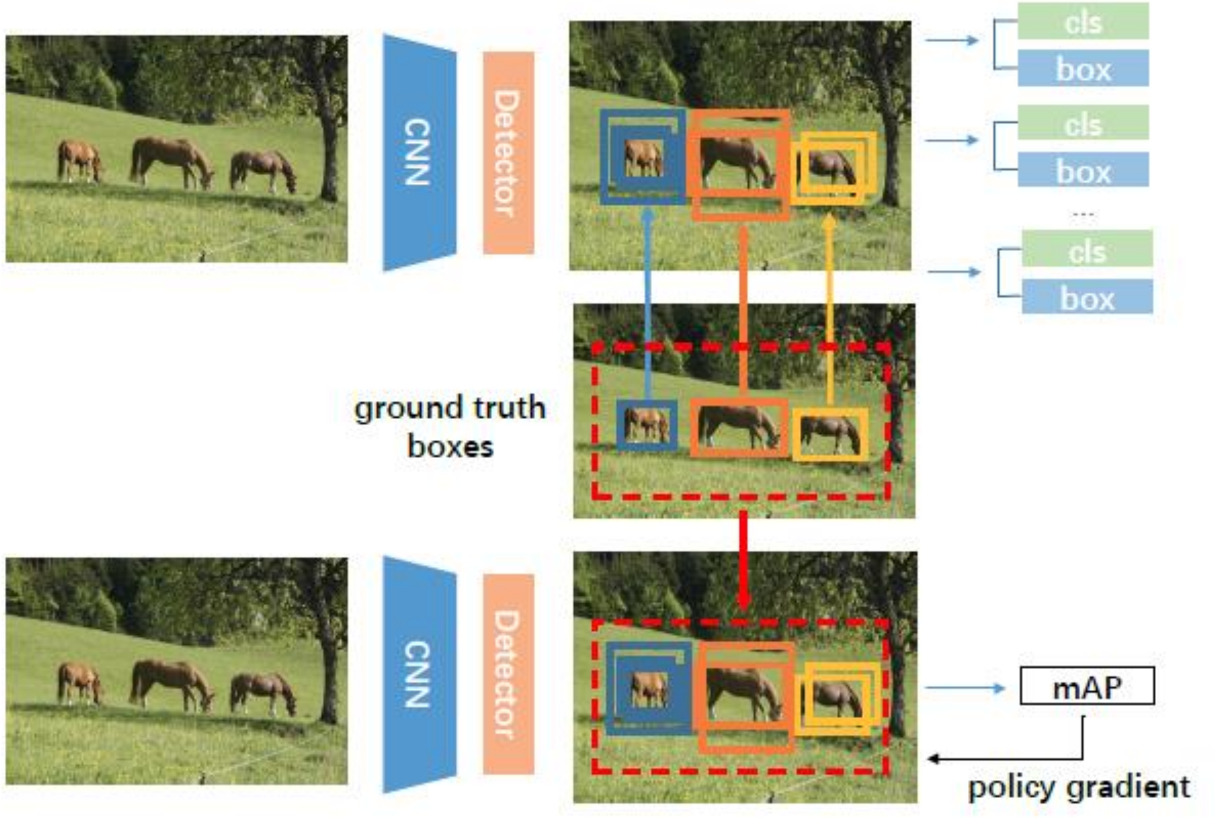
A brief introduction to policy gradient theorem

- Code (pytorch):

```python
def update_policy(states, actions, returns):
  action_probs = policy(states)
  action_dist = torch.distributions.Categorical(action_probs)
  action_loss = -action_dist.log_prob(actions) * returns
  entropy = action_dist.entropy()
  loss = torch.mean(action_loss - 1e-4 * entropy)
  optimizer.zero_grad()
  loss.backward()
  optimizer.step()
  return
```

Example: Object Detection

- *The relation information between RoIs that is ignored in Faster R-CNN can be further utilized to improve object detectors.*



Rao, Y., Lin, D., Lu, J., & Zhou, J. Learning Globally Optimized Object Detector via Policy Gradient. CVPR2018.

Example: Object Detection

- The objective of object detection can be formulated as:

$$\mathcal{H} = \max_\theta \mathrm{mAP}(F_\theta(I), B),$$
$$\text{subject to} \quad |F_\theta(I)| \leq N_{bb}$$

- We can use a policy gradient method to compute the expected gradient of the non-differentiable reward function as follows:

$$\nabla L_I(\theta) = -\mathbb{E}_a[r(a)\nabla_\theta \log(p_a)] \qquad (6)$$

- Action a can be defined as selecting a set of bounding boxes from all candidates

Rao, Y., Lin, D., Lu, J., & Zhou, J. Learning Globally Optimized Object Detector via Policy Gradient. CVPR2018.

Example: Object Detection

p_a is the probability of action a, therefore, $p_a = \prod_{b\in a} p_b$. We can further simplify Equation 6 as:

$$\begin{aligned}
&\mathbb{E}_a[r(a)\nabla_\theta \log(p_a)] \\
= &\sum_a p(a)r(a)\nabla_\theta \log(\prod_{b\in a} p_b) \\
= &\sum_a [p(a)r(a) \sum_{b\in B'} [\delta(a,b)\nabla_\theta \log(p_b)]] \\
= &\sum_{b\in B'} [\nabla_\theta \log(p_b) \sum_a [p(a)r(a)\delta(a,b)]]
\end{aligned}$$

define $r(b) = \sum_a [p(a)r(a)\delta(a,b)]$.

- Sampling several actions during a single gradient calculation is more efficient.

Rao, Y., Lin, D., Lu, J., & Zhou, J. Learning Globally Optimized Object Detector via Policy Gradient. CVPR2018.

Example: Object Detection

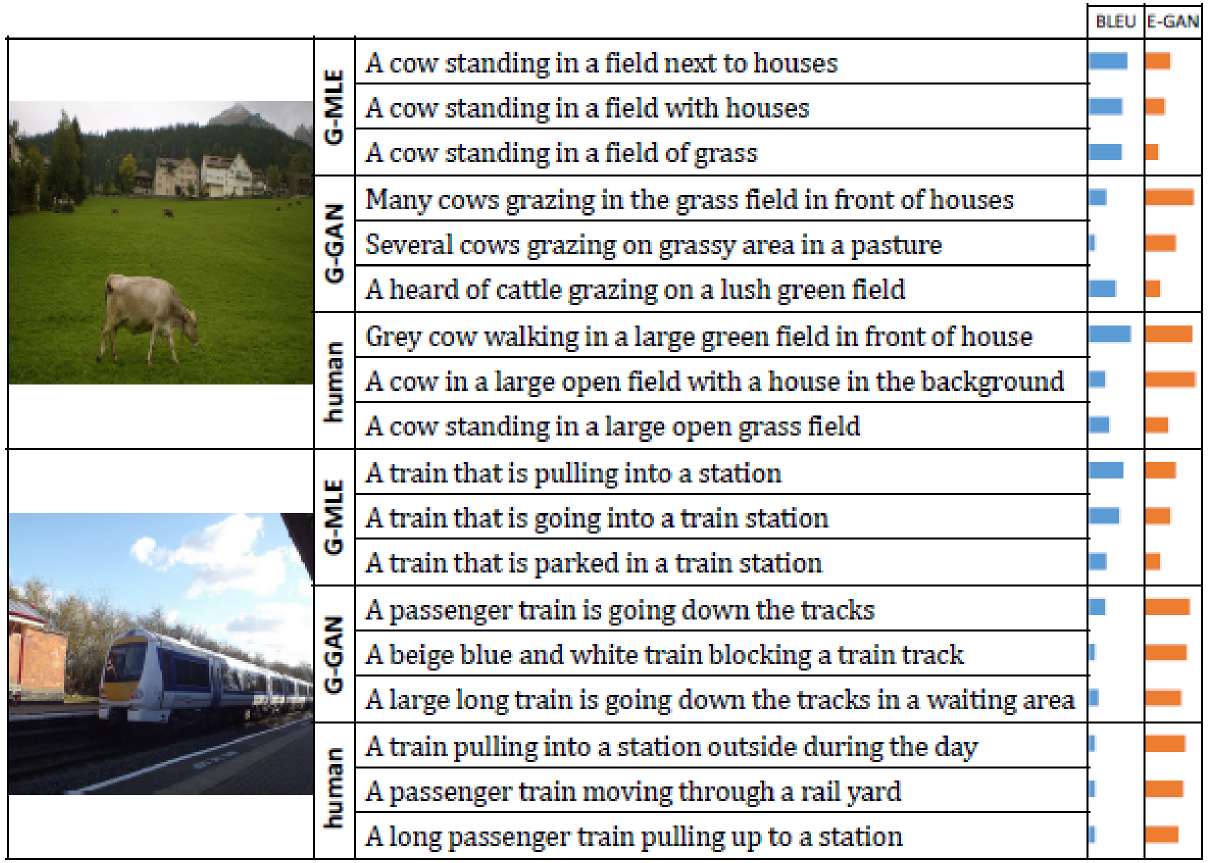
- Experiment results:

Detection model	training method	greedy NMS	soft NMS	mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L
Faster R-CNN	standard	✓		36.3	57.3	38.8	17.7	42.4	51.4
Faster R-CNN	standard		✓	36.9	57.2	40.1	18.0	42.7	52.1
Faster R-CNN	OHEM	✓		36.9	57.3	40.2	17.7	42.7	52.4
Faster R-CNN	ours ($\gamma = 0$)	✓		37.6	60.0	40.2	19.6	42.6	52.0
Faster R-CNN	ours ($\gamma = 1$)	✓		38.3	60.6	40.9	20.7	43.2	52.6
Faster R-CNN	ours ($\gamma = 1$)		✓	**38.5**	**60.8**	**41.3**	**20.9**	**43.4**	**52.7**
Faster R-CNN with FPN	standard	✓		37.7	58.5	40.8	19.3	41.7	**52.3**
Faster R-CNN with FPN	ours ($\gamma = 1$)	✓		**39.5**	**60.2**	**43.3**	**22.7**	**44.1**	51.9

Rao, Y., Lin, D., Lu, J., & Zhou, J. Learning Globally Optimized Object Detector via Policy Gradient. CVPR2018.

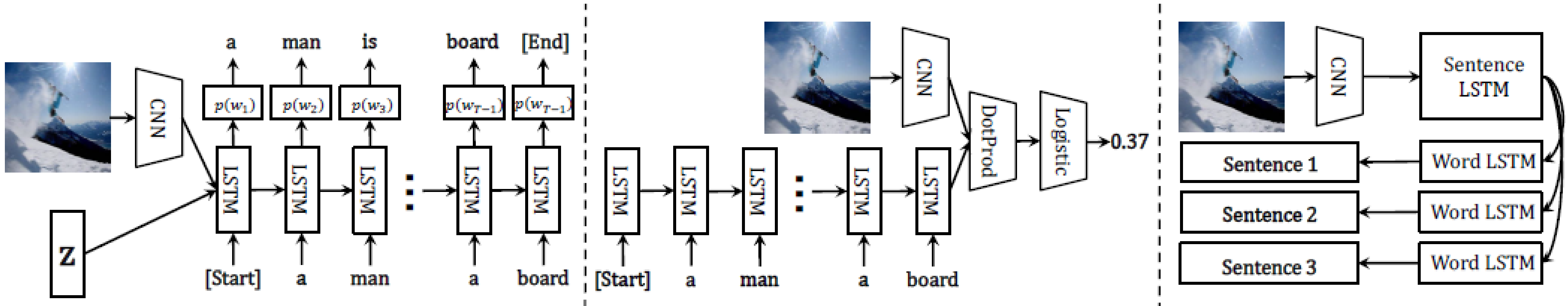
Example: Image Caption

- *Existing efforts primarily focus on fidelity, while other essential qualities of human languages, e.g. naturalness and diversity, have received less attention.*



Dai, B., Fidler, S., Urtasun, R., & Lin, D. Towards Diverse and Natural Image Descriptions via a Conditional GAN. ICCV2017

Example: Image Caption



(a) G for sentence generation (b) E for sentence generation (c) G for paragraph generation

- G: a generator to produce descriptions conditioned on images
- E: an evaluator to assess how well a description fits the visual content.

Dai, B., Fidler, S., Urtasun, R., & Lin, D. Towards Diverse and Natural Image Descriptions via a Conditional GAN. ICCV2017

Example: Image Caption

- Difficulties:
  - The production of sentences is a discrete sampling process, which is non-differentiable.
  - A sentence can only be evaluated when it is completely generated.

- Solutions:
  - Use policy gradient to compute gradient
  - Evaluate an expected future reward when the sentence is partially generated

$$V_{\boldsymbol{\theta},\boldsymbol{\eta}}(I, \mathbf{z}, S_{1:t}) = \mathbb{E}_{S_{t+1:T} \sim G_{\boldsymbol{\theta}}(I,\mathbf{z})}[r_{\boldsymbol{\eta}}(I, S_{1:t} \oplus S_{t+1:T})].$$

  - we can derive the gradient of this objective w.r.t. θ as:

$$\tilde{\mathbb{E}}\left[\sum_{t=1}^{T_{\max}} \sum_{w_t \in \mathcal{V}} \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(w_t | I, \mathbf{z}, S_{1:t-1}) \cdot V_{\boldsymbol{\theta}',\boldsymbol{\psi}}(I, \mathbf{z}, S_{1:t} \oplus w_t)\right]$$

Dai, B., Fidler, S., Urtasun, R., & Lin, D. Towards Diverse and Natural Image Descriptions via a Conditional GAN. ICCV2017

Example: Image Caption

- Experiment results:

		BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr	SPICE	E-NGAN	E-GAN
COCO	human	0.290	0.192	0.240	0.465	0.849	**0.211**	0.527	**0.626**
	G-MLE	**0.393**	**0.299**	**0.248**	**0.527**	**1.020**	0.199	0.464	0.427
	G-GAN	0.305	0.207	0.224	0.475	0.795	0.182	**0.528**	0.602
Flickr	human	0.269	0.185	0.194	0.423	0.627	0.159	0.482	**0.464**
	G-MLE	**0.372**	**0.305**	**0.215**	**0.479**	**0.767**	**0.168**	0.465	0.439
	G-GAN	0.153	0.088	0.132	0.330	0.202	0.087	**0.582**	0.456

Table 1: This table lists the performances of different generators on MSCOCO and Flickr30k. On BLEU-{3,4}, METEOR, ROUGE_L, CIDEr, and SPICE, *G-MLE* is shown to be the best among all generators, surpassing human by a significant margin. While *E-NGAN* regard *G-GAN* as the best generator, *E-GAN* regard *human* as the best one.

Dai, B., Fidler, S., Urtasun, R., & Lin, D. Towards Diverse and Natural Image Descriptions via a Conditional GAN. ICCV2017

Example: Image Caption

- Experiment results:

$\mathbf{z}_1$	a baseball player holds a bat up to hit the ball	a man riding a snowboard down a slope	a group of people sitting around a table having a meal in a restaurant	a group of men dressed in suits posing for a photo
$\mathbf{z}_2$	a baseball player holding white bat and wear blue baseball uniform	a person standing on a snowboard sliding down a hill	a young man sitting at a table with coffee and a lot of food	a couple of men standing next to each other wearing glasses
$\mathbf{z}_3$	a professional baseball player holds up his bat as he watches	a man is jumping over a snow covered hill	a pretty young man sitting next to two men in lots of people	some people dressed in costume and cups

Figure 6: This figure shows example images with descriptions generated by *G-GAN* with different $\mathbf{z}$.

Dai, B., Fidler, S., Urtasun, R., & Lin, D. Towards Diverse and Natural Image Descriptions via a Conditional GAN. ICCV2017