EN. 520.650 Final exam

Qihua Gong

Q 1. a)

4	0	0	36

$V_1(s)$ is equal to optimal one-step rewards from each state, it gives reward equal to the number on each square.

b)

6	1	9	54

first square: $4 + \gamma \cdot 4 = 4 + 2 = 6$

last square: $36 + \gamma \cdot 36 = 36 + 18 = 54$

In optimal policy for a two-step horizon is to move outward the closer side.

so, for the third square: $\frac{1}{2} \times 0 + \frac{1}{2} \times 18 = 9$

second square $= \frac{1}{4} \times 4 = 1$

c)

8	8	24	72

last square: $\Sigma_{j=0}^{\infty} (\gamma^j \cdot 36) = 36 \cdot (1 + \frac{1}{2} + \frac{1}{4} + \cdots)$

$= 36 \times 2 = 72$

$\because V^*(MR) = \frac{1}{2}(0 + \gamma V^*(MR)) + \frac{1}{2}(0 + \gamma V^*(R)) = \frac{1}{4} V^*(MR) + \frac{1}{4} V^*(R)$

$V^*(MR) = \frac{1}{3} V^*(R)$

$\therefore$ third square $= \frac{1}{3} \times 72 = 24$

first left square $= 2 \times 4 = 8$

second square $= 24 \div 3 = 8$

Q2. a) $\forall l \in \{1, \dots, \mathcal{L}\}$, $a_l = W_l \cdot O_{l-1} + b_l$

$\forall l \in \{1, \dots, \mathcal{L}\}$, $O_l = \sigma_l(a_l)$

$f = \sigma_{\mathcal{L}}(W_{\mathcal{L}} \cdot \sigma_{\mathcal{L}-1}(W_{\mathcal{L}-1} \cdots \sigma_2(W_2 \cdot \sigma_1(W_1 x + b_1) + b_2) \cdots + b_{\mathcal{L}-1}) + b_{\mathcal{L}})$

b) ReLU will take $\mathcal{L}$ additions, $\mathcal{L}$ multiplications and $\mathcal{L}$ nonlinearities. Neither will dominate.

c) backpropogation rule: $\frac{\partial f}{\partial a_l} = ReLU'(a_l) \cdot \frac{\partial f}{\partial o_l}$, $\frac{\partial f}{\partial a_{l-1}} = W_l^T \cdot \frac{\partial f}{\partial a_l}$

$\nabla W_l f = \frac{\partial f}{\partial a_l} \cdot O_{l-1}^T$, $\nabla b_l f = \frac{\partial f}{\partial a_l}$

d) To update ~~the~~ all the parameters in the network, we need $3\mathcal{L}$ multiplications and $\mathcal{L}$ nonlinearities of ReLU operations. So, in the backpropagation, the matrix multiplies dominate.

Q3. a) True. The mini batch gradient descent uses an empirical estimate of gradient from a small batch, so the examples in a batch should be iid. Only when the model is correlated, the estimate will become biased and the model will fail to learn.

b) If first training on dataset 1 and then go to dataset 2, the model will forget what it learned from the positive examples and will always predict the negative one.

c) 1) computational tractability, It can reduce the compute in the netward and ehance the network performance.

2) It can reduce over fitting and make translation invariant.

d) It will cause all predictions to be positive

e) No. The searching is too crowed between $0.1 \sim 1$, We are searching between $0.01 \sim 1$, the five search only one is in range $0.01 \sim 0.1$

f) i) It can have less computationally expensive ~~and~~ and faster performance.

ii) It can have faster convergence and less divergence

iii) Also, it can have faster convergence and stable learning, more efficiency

Q4. In GANs, the mode collapse occurs when the diversity of ~~gen~~ generated samples become less than that of the real data. Also means $\min_G \max_D V(G,D) \neq \max_D \min_G V(G,D)$.

Also, the model collapse may happens ~~when~~ with the training objective formulations, low capacity generators or weak discriminator functions.

One of the way to avoid the collapse is to use discriminator augmentation, which modifies the discriminator to make decision based on multiple samples of real or generated distributions. Also, we can use mutual information to mitigate the collapse by increasing the entropy of the generated samples.

Q5. a) (i) True

(ii) False. fgsm is not iterative

(iii) False. dropout doesn't work in test time

(iv) True

b) $x^* = x + \varepsilon \operatorname{sign}(\frac{\partial J}{\partial x}) = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + 0.01 \times \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.01 \\ 1.99 \\ 3.01 \end{bmatrix}$

© Disagree. For example: $\hat{y} = \sigma(w x + b)$

$x^* = x + \varepsilon \cdot \operatorname{sign}(w^T)$

output: $\hat{y}^* = \sigma(w(x + \varepsilon \cdot \operatorname{sign}(w^T)) + b) = \sigma(wx + b + \varepsilon \|w\|_1)$

If x and w have dimension n and the average magnitude of elements w is m, so the change input function grows as $\theta(\epsilon mn)$. x and ϵ grows linearly. As a result, even ϵ is small, we can still perturb the elements of the input by ϵ and achieve a large deviation in the output when x is large.

c) choose B: non-saturating cost

Because the gradient towards the begining of training is large.

d) Because the losses are different to quality models. The loss in epoch 1~100 are respect to a discriminator ~~that~~ ~~is~~ will significantly improve and the loss of the discriminator also improve.

Q6. a) There's a famous example in AI bias about the Face Recognition Bias.

In the FR bias, there's many example of gender bias and skinstone bias.

There's a news about the gender and racial bias found in Amazon's facial ~~~~ recognition technology. The research shows the divergent error rates across demographic groups, with the poorest accuracy consistently found in subjects who are female, black and 18~30 years old.

For the way of solution, there's a Distill and De-bias way we talked in class. It can mitigating Bias in FR using knowledge distillation. FR network attend to different spatial regions, depending on demographic groups. But in the DD method, we need to put the attention on the dissimiliar attention regions for male and female and the regions for dark the light skin. Use the average attention map to train a specific model and then reduce gender and racial bias and maintain high verification performance. D&D can be used for producing de-biased descriptors in ~~~~ applications that do not require preserving privacy.

b) The Bayesian model is a kind of interpretable AI. It's a probabilistic graphical version that represent a hard and rapid of variables and their conditional dependencies through a direct acyclic graph. It can use to encompass prediction, anomaly detection, diagnostics. For the reason bayesian is a step by step probabilistic model, so it's compact, flexible and iterpretable illustration of a joint possibility distribution. We can check the causal relationship between variables. The Bayesian network has the advantage of being easy and mathematically consistent across multiple accuracy records and surprising assets. It's as good as the reliability of its initial information.

c) CAM is a explanation method used for CNN. In the network, the stack of fully connected layers at the very end of the model has been replaced by layer GAP. It averages the activations of each feature map and concatenates these averages and output them as a vector. Then a weight sum of the vector is fed to the final softmax loss layers. So, we weight the features maps in CAM using weights from the network last fully connected layers. Grad-CAM weights the feature maps using α values calculate from gradients. The Grad CAM does not require a particular CNN architecture. GrAd-CAM is a generalization of CAM, a method that does not require using a particular architecture. The CAM requires an architecture to produce the prediction.