CS7150.37338.202130 Day 14 Preparation Questions

Aishwarya Vantipuli, Harshita Ved, Aveek Choudhury

TOTAL POINTS

5 / 5

QUESTION 1

1 Question 2 1 / 1

✓ **- 0 pts** **Correct**

- 1 pts Not correct

QUESTION 2

2 Question 3 1 / 1

✓ **- 0 pts** **Correct**

- 1 pts Click here to replace this description.

QUESTION 3

3 Question 4 1 / 1

✓ **- 0 pts** **Correct**

- 0.5 pts Click here to replace this description.

QUESTION 4

4 Question 5 1 / 1

✓ **- 0 pts** **Correct**

QUESTION 5

5 Question 6 1 / 1

✓ **- 0 pts** **Correct**

- 0.5 pts Click here to replace this description.

Question 2. *What is the process for computing an adversarial example using the fast gradient sign method? Be clear to specify what the inputs and output of this process are.*

Response: The process for computing an adversarial example using the fast gradient sign method can be represented as -

$$\eta = \epsilon.sign(\Delta_x J(\theta, x, y))$$

The output: x^* which is the new example that is calculated to fool the given network is:

$$x^* = x + \eta$$

where -
θ represents the parameters of a model

x is the input given to the model which is a training example

y is the target associated with x which is the class it belongs to.

$J(\theta, x, y)$ computes the cost used to train the network or the loss

ϵ is the small error term, i.e. a user-specified hyper-parameter giving the amount user wants to perturb the input.

η represents the computed value of optimal max-norm constrained perturbation to be applied to a training example to generate the adversarial example.

The adversarial example is created by either adding or subtracting the error ϵ to each pixel. The decision to add or subtract ϵ depends on the sign of the gradient for a pixel.

1 Question 2 1 / 1

✓ **- 0 pts** **Correct**

- 1 pts Not correct

Directions: Read 'Robust Physical-World Attacks on Deep Learning Models'

- Read the whole paper. You can skip Section 4.

Question 3. *Why can't the approach of Goodfellow et al. be directly applied to generate a physical attack on a real Stop sign?*

Response: Goodfellow et al. approach is more on digital perturbation of input images, while to generate a physical attack on a real stop sign requires physical perturbations to input object. Challenges with generating robust physical perturbations Goodfellow paper didn't address:
a) Background changes
b) Environmental conditions
c) Physical limits on fabrication
d) Limits on perceptibility

The main challenge with generating robust physical perturbations which is not addressed in Goodfellow paper is environmental variability. Cyber-physical systems operate in noisy physical environments that can destroy perturbations created using current digital-only algorithms (Fast Gradient Method by Goodfellow et al). Hence authors proposed RP2 method which examines physical perturbations on real objects (stop sign) under varying conditions like changes in lighting/weather conditions, and the presence of debris on the camera or on the road sign.

Question 4. *In modeling environmental conditions, the authors collected some real images and made synthetic transformations. What data did they collect? What synthetic transformations did they make? Why did they do this?*

Response:
What data did they collect?:
Authors choose road sign classification as target domain and collected various images containing stop signs.

What synthetic transformations did they make?:
Authors model a space of (physical and digital) transformations, which involves taking images of the real physical target object from several angles, distances, and lighting conditions. For synthetic variations, they randomly cropped the object within the image, change the brightness, and add spatial transformations to simulate other possible condition. They masked images to project the computed perturbations to a physical region on the surface of the object.

Why did they do this?:
Authors goal is to examine whether it is possible to create robust physical perturbations for real-world objects that mislead classifiers to make incorrect predictions even when images are taken in a range of varying physical conditions including Environmental Conditions, Spatial Constraints, Physical Limits on Imperceptibility and Fabrication Error. They want to introduce bias by simulating real world scenario.

2 Question 3 1 / 1

✓ **- 0 pts** **Correct**

- 1 pts Click here to replace this description.

Directions: Read 'Robust Physical-World Attacks on Deep Learning Models'

- Read the whole paper. You can skip Section 4.

Question 3. *Why can't the approach of Goodfellow et al. be directly applied to generate a physical attack on a real Stop sign?*

Response: Goodfellow et al. approach is more on digital perturbation of input images, while to generate a physical attack on a real stop sign requires physical perturbations to input object. Challenges with generating robust physical perturbations Goodfellow paper didn't address:
a) Background changes
b) Environmental conditions
c) Physical limits on fabrication
d) Limits on perceptibility

The main challenge with generating robust physical perturbations which is not addressed in Goodfellow paper is environmental variability. Cyber-physical systems operate in noisy physical environments that can destroy perturbations created using current digital-only algorithms (Fast Gradient Method by Goodfellow et al). Hence authors proposed RP2 method which examines physical perturbations on real objects (stop sign) under varying conditions like changes in lighting/weather conditions, and the presence of debris on the camera or on the road sign.

Question 4. *In modeling environmental conditions, the authors collected some real images and made synthetic transformations. What data did they collect? What synthetic transformations did they make? Why did they do this?*

Response:
What data did they collect?:
Authors choose road sign classification as target domain and collected various images containing stop signs.

What synthetic transformations did they make?:
Authors model a space of (physical and digital) transformations, which involves taking images of the real physical target object from several angles, distances, and lighting conditions. For synthetic variations, they randomly cropped the object within the image, change the brightness, and add spatial transformations to simulate other possible condition. They masked images to project the computed perturbations to a physical region on the surface of the object.

Why did they do this?:
Authors goal is to examine whether it is possible to create robust physical perturbations for real-world objects that mislead classifiers to make incorrect predictions even when images are taken in a range of varying physical conditions including Environmental Conditions, Spatial Constraints, Physical Limits on Imperceptibility and Fabrication Error. They want to introduce bias by simulating real world scenario.

3 Question 4 1 / 1

✓ **- 0 pts** **Correct**

- 0.5 pts Click here to replace this description.

Question 5. *Explain the meaning and purpose of each term of equation (3).*

Response: The equation (3) is for the final robust spatially- constrained perturbation optimization, mentioned below:

$$\begin{aligned}\underset{\delta}{\operatorname{argmin}}\ & \lambda||M_x \cdot \delta||_p + NPS \\ & + \mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*)\end{aligned}$$

where meaning and purpose of each term is described below:
a) x_i = single image drawn from X^V.

b) δ = perturbation to be added to the input x, such that the perturbed instance $x' = x + \delta$ is misclassified by the target classifier f_θ $(\cdot)$

c) $y*$ is the target class

d) $J(\cdot,\cdot)$ is the loss function, which measures the difference between the model's prediction (f_θ $(x + \delta)$) and the target label $y*$. Solving for this optimization would result in calculating gradient Δ_x $J(f_\theta(x + \delta),y*)$

e) λ is a hyper-parameter that controls the regularization of the distortion.

f) M_x is a perturbation mask i.e. a matrix of 0s and 1s whose dimensions are the same as the size of input to the road sign classifier and is determining where on the road sign a perturbation is used.

g) Non-Printability Score (NPS) - To account for fabrication error, an additional term is added to our objective function that models printer color reproduction errors. Given a set of printable colors (RGB triples) P and a set R() of (unique) RGB triples used in the perturbation that need to be printed out in physical world, the NPS is given by:

$$NPS = \sum_{\hat{p} \in R(\delta)} \prod_{p' \in P} |\hat{p} - p'|$$

h) $T_i(\cdot)$ = denotes the alignment function that maps transformations on the object to transformations on the perturbation (e.g. if the object is rotated, the perturbation is rotated as well).

i) X^V = the space of transformed inputs, so x_i has already been transformed.

j) f_θ - The network function.

k) $\mathbb{E}$ - The expected value over the random variable X^V for calculating the average cost over all of the images and transformations.

Now equation (3) is made of addition of three terms:

1. $\lambda||M_x \cdot \delta||_p$ = A regularization term, with · equals regularization hyper-parameter.
It exists for not allowing any singular weight to become too extreme. Here, the mask is element-wise multiplied with delta, and then a scalar is produced by taking the Lp norm of this matrix and multiplying by ·

2. Non-Printability Score (NPS) = same as mentioned above

3. $\mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*)$ equals cost term between the new prediction of network and the target class. This is to force the model to predict the target class with as high accuracy as possible under perturbation.

Question 6. *How did the authors ensure that the adversarial perturbation is restricted to the area of the Stop sign (and not the background)? How did they ensure that the perturbation only takes up a small fraction of the Stop sign's area?*

Response: To ensure that the adversarial perturbations are only applied to the surface area of the target object *o* i.e. stop sign here (considering the spatial constraints and physical limits on imperceptibility), authors introduce a perturbation mask.

This mask serves to project the computed perturbations to a physical region on the surface of the object (i.e. road sign). In addition to providing spatial locality, the mask also helps generate perturbations that are visible but inconspicuous to human observers. To do this, an attacker can shape the mask to look like graffiti—commonplace vandalism on the street that most humans expect and ignore, therefore hiding the perturbations "in the human psyche."

The perturbation mask is a matrix M_x whose dimensions are the same as the size of input to the road sign classifier. M_x contains zeroes in regions where no perturbation is added, and ones in regions where the perturbation is added during optimization.

To ensure that the perturbation only takes up a small fraction of the Stop sign's area perturbation are computed using L_1 regularization and mask occupying entire stop sign, then the results help humans as reference to place the mask and finally perturbation are recomputed using L_2 with stricter mask than what is obtained above.

4 Question 5 1 / 1

✓ - 0 pts Correct

Now equation (3) is made of addition of three terms:

1. $\lambda||M_x \cdot \delta||_p$ = A regularization term, with · equals regularization hyper-parameter.
It exists for not allowing any singular weight to become too extreme. Here, the mask is element-wise multiplied with delta, and then a scalar is produced by taking the Lp norm of this matrix and multiplying by ·

2. Non-Printability Score (NPS) = same as mentioned above

3. $\mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*)$ equals cost term between the new prediction of network and the target class. This is to force the model to predict the target class with as high accuracy as possible under perturbation.

Question 6. *How did the authors ensure that the adversarial perturbation is restricted to the area of the Stop sign (and not the background)? How did they ensure that the perturbation only takes up a small fraction of the Stop sign's area?*

Response: To ensure that the adversarial perturbations are only applied to the surface area of the target object *o* i.e. stop sign here (considering the spatial constraints and physical limits on imperceptibility), authors introduce a perturbation mask.

This mask serves to project the computed perturbations to a physical region on the surface of the object (i.e. road sign). In addition to providing spatial locality, the mask also helps generate perturbations that are visible but inconspicuous to human observers. To do this, an attacker can shape the mask to look like graffiti—commonplace vandalism on the street that most humans expect and ignore, therefore hiding the perturbations "in the human psyche."

The perturbation mask is a matrix M_x whose dimensions are the same as the size of input to the road sign classifier. M_x contains zeroes in regions where no perturbation is added, and ones in regions where the perturbation is added during optimization.

To ensure that the perturbation only takes up a small fraction of the Stop sign's area perturbation are computed using L_1 regularization and mask occupying entire stop sign, then the results help humans as reference to place the mask and finally perturbation are recomputed using L_2 with stricter mask than what is obtained above.

5 Question 6 1 / 1

✓ **- 0 pts** **Correct**

- 0.5 pts Click here to replace this description.