تمرین سری دوم SPML ❀ امیرمحمد غزنی 402212269

سوال ①

الف)

MobileNet v1

Depth wise Conv

$$3 \times 3 \times \underbrace{3}_{\#\text{Channels}} + \underbrace{3}_{\text{biases}} = 30$$

Point wise Conv

$$1 \times 1 \times \underbrace{3}_{\#\text{Channels}} \times \underbrace{128}_{\#\text{filters}} + \underbrace{128}_{\text{biases}} = 512$$

Sum of parameters = 542

CNN

Convolution

$$3 \times 3 \times \underbrace{3}_{\#\text{Channels}} \times \underbrace{128}_{\#\text{filters}} + 128 = 3584$$

ب)

18 point wise

$$1 \times 1 \times 3 \times 18 + 18 = 72$$

Depth wise

$$3 \times 3 \times 18 + 18 = 180$$

ادامه ب)

point wise

$1 \times 1 \times 18 \times 3 + 3 = 57$

Sum of all $= 72 + 180 + 57 = 409$

ج) بدلیل اینکه CNN معمولی تعداد پارامترهای خیلی بیشتری دارد و باعث تولید redundancy برای مدل می‌شود پس بیشتر به perturbation های جزئی مقاوم‌تر است. پس بطور کلی در برابر حمله مقاوم‌تری باشد.

اما از سوی دیگر نشان کنیم، با افزایش پارامترهای مدل، decision boundary مدل ما بسیار پیچیده‌تری می‌شود و ممکن است مرزهای بسیار دقیق حول و اطراف داده‌ها در نظر بگیرد و نسبت به آشفتگی‌های جزئی حساس‌تر شده و باعث دسته‌بندی اشتباه شود پس در این صورت CNN معمولی چون پارامترهای بیشتری دارد از هم ضعیف‌تر است.

سوال ②

الف)

Untargeted attack

$$\max_{\|\delta\| \leq \varepsilon} \ell(h(x+\delta), y)$$

$$x' = x + \varepsilon \, \text{Sign}\left(\nabla_x \ell(h(x), y)\right)$$

با توجه به اینکه در حالت untargeted به دنبال این موضوع هستیم صرفا پیشبینی Classifier را هر لیبلی به جز y تبدیل کنیم پس فقط کافیست با perturbation کمی داده را از مرزبندی کلاس y خارج کنیم. امّا در حالت targeted چون می‌خواهیم پیشبینی به یک کلاس خاص تبدیل شود ممکن است نیاز به perturbation بزرگتر و حمله پرچالش‌تری داشته باشیم.

Targeted attack

$$\min_{\|\delta\| \leq \varepsilon} \ell(h(x+\delta), y')$$

$$x' = x - \varepsilon \, \text{Sign}\left(\nabla_x \ell(h(x), y')\right)$$

پس حمله untargeted نسبتا آسان‌تر خواهد بود و موفقیت بیشتری دارد.

سوال ②

ب) اگر $h(x)$ خطی باشد: $h(x) = w^T x + b$

$$\max \; \ell\left(h(x+\delta), y\right)$$
$$\|\delta\|_p \leq \varepsilon$$

اکنون لاگرانژ را بنویسیم:

$$\begin{cases} f(\delta) = \ell\left(h(x+\delta), y\right) \\ g(\delta) = \|\delta\|_\infty - \varepsilon \leq 0 \end{cases}$$

$$L(\delta, \alpha) = \ell\left(h(x+\delta), y\right) + \alpha\left(\|\delta\|_\infty - \varepsilon\right)$$

$$= \ell\left(w^T(x+\delta), y\right) + \sum_{i=1}^{n} \alpha_i \left(|\delta_i| - \varepsilon\right)$$

KKT Conditions

1. Stationary (maximizing): $-\nabla_\delta \ell\left(w^T(x+\delta), y\right) + \nabla_\delta \alpha\left(\|\delta\|_\infty - \varepsilon\right) = 0$
2. Primal feasibility: $\|\delta\|_\infty \leq \varepsilon \Rightarrow \|\delta\|_\infty - \varepsilon \leq 0$
3. Dual feasibility: $\alpha_i \geq 0$ for $i = 1, 2, \ldots, n$
4. Complementary slackness: $\sum_{i=1}^{n} \alpha_i \left(|\delta_i| - \varepsilon\right) = 0$

$$\text{for each } i \Rightarrow \alpha_i \left(|\delta_i| - \varepsilon\right) = 0$$

$$-\nabla_{\delta_i} \ell\left(w^T(x+\delta), y\right) + \nabla_{\delta_i} \sum \alpha_i \left(|\delta_i| - \varepsilon\right) = 0 \quad (2$$

$$-\left(\nabla_h \ell \cdot \nabla_{\delta_i} h\right) + \alpha_i \, sign(\delta_i) = 0 \Rightarrow$$

$$-\left(\nabla_h \ell \cdot w_i\right) + \alpha_i \, sign(\delta_i) = 0 \Rightarrow$$

$$\alpha_i = \frac{\nabla_h \ell \cdot w_i}{sign(\delta_i)} \qquad (\delta_i \neq 0)$$

(>

$$\delta_i = sign(\delta_i) \, |\delta_i|$$

Complementary slackness

$$\sum_{i=1}^{n} \alpha_i \left(|\delta_i| - \varepsilon\right) = 0 \xrightarrow{\alpha_i \geq 0} \alpha_i \left(|\delta_i| - \varepsilon\right) = 0$$

$$\alpha_i \left(\frac{\delta_i}{sign(\delta_i)} - \varepsilon\right) = 0 \Rightarrow \alpha_i = 0 \ (1) \Rightarrow \nabla_{\delta_i} \ell = 0$$

یعنی: مقدار بهینه δ_i رسیدیم زمانی که $|\delta_i| < \varepsilon$ بود.

$$\frac{\nabla_h \ell \cdot w_i}{sign(\delta_i)} \cdot \underbrace{\frac{\delta_i}{sign(\delta_i)}}_{+1} - \frac{\nabla_h \ell \, w_i}{sign(\delta_i)} \varepsilon = 0 \Rightarrow$$

$$\delta_i = \frac{\varepsilon}{sign(\delta_i)} \ (2) \Rightarrow |\delta_i| = \varepsilon$$

هـ) در عبارت قسمت "ب" داشتیم :

Stationary

$$-\nabla_\delta \ell\left(h(x), y\right) + \nabla_\delta \alpha\left(\|\delta\|_\infty - \varepsilon\right) = 0$$

$$\nabla_\delta \ell\left(h(x), y\right) = \nabla_\delta \alpha\left(\|\delta\|_\infty - \varepsilon\right)$$

به ازای دو مقدار بدست آمده در بخش "د" یعنی $\alpha_i = 0$ یا $|\delta_i| = \varepsilon$ با جایگذاری در عبارت بالا می‌توانیم نتیجه بگیریم در حالت بهینه قرار داریم.

$$\|\delta\|_\infty = \max_i |\delta_i|$$

چون به ازای هر i داشتیم :

$$|\delta_i| = \varepsilon \Rightarrow \nabla_{\delta_i} \alpha_i (\varepsilon - \varepsilon) = 0 \Rightarrow \boxed{\nabla_\delta \ell = 0}$$

or

$$\alpha_i = \varepsilon \Rightarrow \nabla_{\delta_i} 0\left(|\delta_i| - \varepsilon\right) = 0 \Rightarrow \boxed{\nabla_\delta \ell = 0}$$

← پس با هر دو مقدار بدست آمده در بخش "د" می‌توانیم بگوییم مقدار مشتق Loss نسبت به δ صفر شده در نتیجه در نقطه بهینه یا ماکزیمم تابع قرار داریم.

سوال ③

الف) ابتدا یک perturbation اولیه انتخاب می‌شود. مانند v که می‌تواند صفر مقداردهی شود.
سپس بصورت iterative روی تمام data point ها v را اعمال می‌کنیم به امید آنکه کلاس آن داده تغییر کند. اگر v نتوانست کلاس را تغییر بدهد.
سپس یک Δv ای پیدا می‌کنیم که اگر به داده perturb شده اضافه شود بتواند کلاس داده را تغییر دهد. سپس $v \leftarrow v + \Delta v$ کرده و سپس v را به حد بالای آن (ϵ) project می‌کنیم تا شرط $\|v\| \leq \epsilon$ را ارضا کنیم.

این کار را بصورت تکرار شونده روی هم داده ها انجام می‌دهیم. اگر $1 - \delta$ از کل داده ها کلاس‌بندی اشتباه شوند کارمان به پایان می‌رسد و اگر این طور نشود دوباره همان حلقه تکرار شونده را برای کل داده‌ها انجام می‌دهیم.

ب)

data point ها در ابعاد بالا فقط بخش کوچکی از فضا را اشغال می‌کنند همچنین مرزهای تصمیم‌گیری بسیار پیچیده بوده و در بسیاری از بعدها بی معنی است.
پس می‌توان با یک آشفتگی Universal در برخی بعدها باعث شد هم داده‌ها شبیه دسته بندی شوند.
این موضوع بخاطر این است که شبکه‌های عمیق صرف الگوهایی را در بخش از فضا یاد می‌گیرند که کلاس ها را از هم جدا کند و یک سری آسیب پذیری مشترک برای هم داده ها در ابعاد بالا بوجود می‌آید که می‌تواند مورد حمله Universal قرار گیرد.

همچنین دلیل دیگری می‌تواند این باشد که شبکه های عمیق در ابعاد بالا روی الگوهای پیچیده overFit می‌شوند که این موضوع منتج به یادگیری ویژگی های غیرمستحکم شده و اغتشاش های Universal ای بوجود می‌آید که مدل ها نسبت به آن آسیب پذیر شود.

ج) برای دیتاست دوبجری two moons سه دایرکشن مختلف برای اضافه کردن نویز را بررسی کردم که به ترتیب در جهت بردارهای $(1,0)$ ، $(0,1)$ و $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ بوده که L_2 norm همگی برابر 1 هستند. در این حالت اگر این بردارها را در ε ضرب کرده و به تصاویر اضافه کنیم می‌توانیم بگوییم نرم میزان تغییرات برابر با ε است.

پس از بررسی متوجه می‌شویم که با افزایش ε در جهت بردارهای $(0,1)$ و $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ حملمان روبه افزایش بوده است اما در جهت بردار $(1,0)$ با افزایش مقدار ε ، تغییرات خاصی در Success rate مشاهده نمی‌شد پس این می‌تواند یک failure seenario برای حملمان باشد. (چارت مربوط به experiment در صفحه بعد قرار داده شده است)

توضیحات اضافه ریاضی برای پارت الف)

$v \leftarrow 0$ initialize

while $\frac{1}{m}\sum_{i=1}^{m} 1_{\hat{k}(x_i+v) \neq \hat{k}(x_i)} \leq 1-\delta$:

به ازای هر داده موجود در دیتاست اگر $\hat{k}(x_i+v) = \hat{k}(x_i)$ است

$$\Delta v_i \leftarrow \arg\min_r \|r\|_2$$

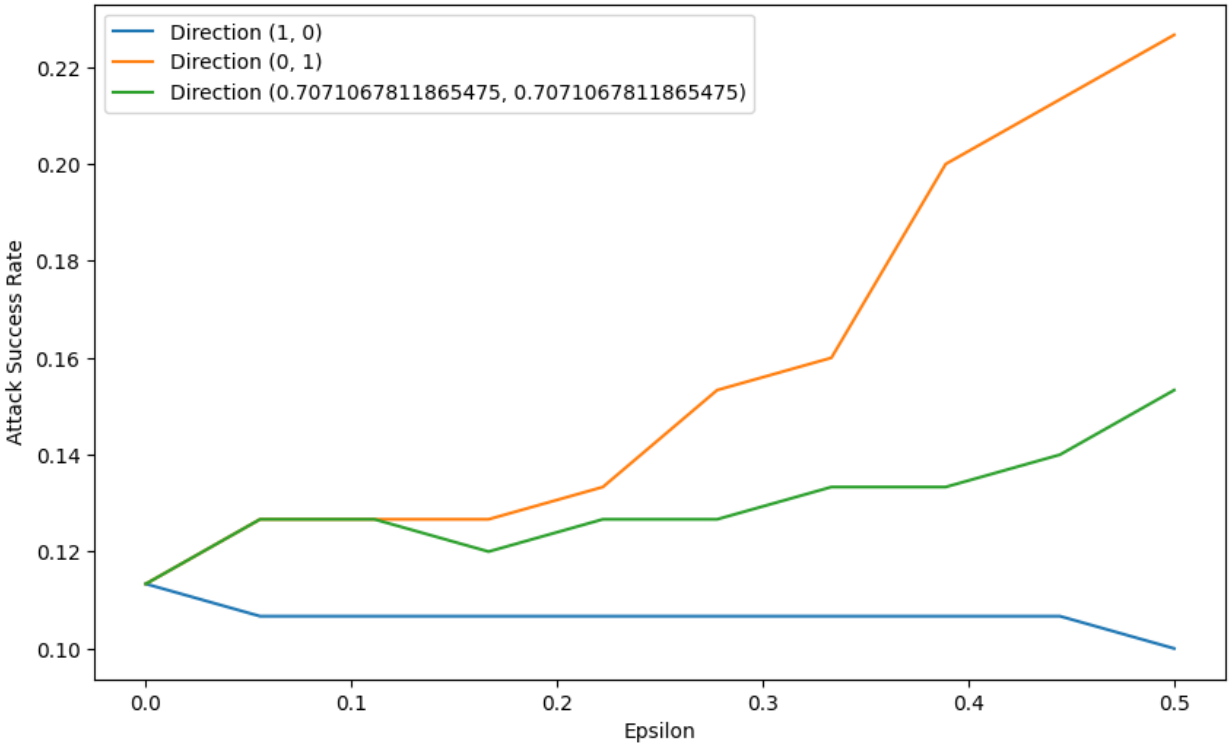
باشرط اینکه $\hat{k}(x_i+v+r) \neq \hat{k}(x_i)$ شود

$$v \leftarrow P_{p,\xi}(v + \Delta v_i)$$

مقدار نهایی perturbation را به مقدار ξ ، project می‌کند.

* این لوپ زمانی که حداقل $1-\delta$ از کل داده‌ها اشتباه دسته‌بندی شوند ادامه می‌یابد.

چارت experiment مربوط به بخش "ج" سوال سوم تئوری:



با توجه به چارت متوجه میشویم که با افزایش epsilon در جهت بردار (1,0) بهبودی در attack success rate مشاهده نمیشود و این یک failure scenario برای حمله‌مان است.