به نام خدا

دانشکده مهندسی کامپیوتر

امنیت و حریم خصوصی در یادگیری ماشین (۴۰۸۱۶)

نیم‌سال اول سال تحصیلی ۱۴۰۴-۱۴۰۳

استاد درس: دکتر امیرمهدی صادق‌زاده

طراحان: سروش وفایی‌تبار، امیرمحمد ایزدی

تمرین دوم

مهلت تحویل: ساعت ۲۳:۵۹ یکشنبه ۱۳ آبان ۱۴۰۳

نکات و قواعد

۱. سوالات خود را زیر پیام مربوطه در Quera مطرح نمایید.

۲. لطفا مطابق تاکید پیشین، حتما آداب‌نامه‌ی انجام تمرین‌های درسی را رعایت نمایید. در صورت تخطی از آیین‌نامه، در بهترین حالت مجبور به حذف درس خواهید شد.

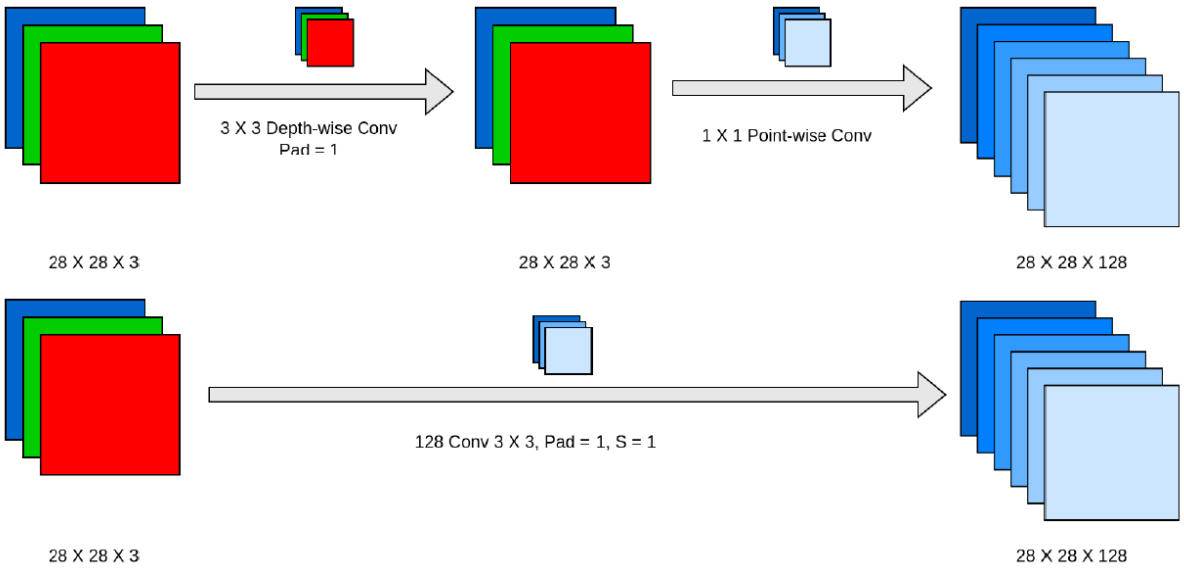
۳. در صورتی که پاسخ‌های سوالات نظری را به صورت دست‌نویس آماده کرده‌اید، لطفا تصاویر واضحی از پاسخ‌های خود ارسال کنید. در صورت ناخوانا بودن پاسخ ارسالی، نمره‌ای به پاسخ ارسال شده تعلق نمی‌گیرد.

۴. همه‌ی فایل‌های مربوط به پاسخ خود را در یک فایل فشرده و با نام SPML_HW۲_StdNum_FirstName_LastName ذخیره کرده و ارسال نمائید.

سوال ۱ کانولوشن عمقی (۲۰ نمره)

یکی دیگر از مشکلاتی که در ارتباط با شبکه‌های ژرف عمیق وجود دارد، تعداد بالای پارامترها و پیچیدگی بالای محاسباتی است. این مشکل استفاده از شبکه‌های CNN بر روی دستگاه‌های کوچک با پردازنده‌های محدود (مانند تلفن همراه) را با دشواری‌هایی همراه می‌سازد. برای حل این مشکل MobileNet است. شبکه MobileNet-V1 از کانولوشن عمقی استفاده می‌کند که تعداد پارامترها و پیچیدگی زمانی را کاهش دهد.

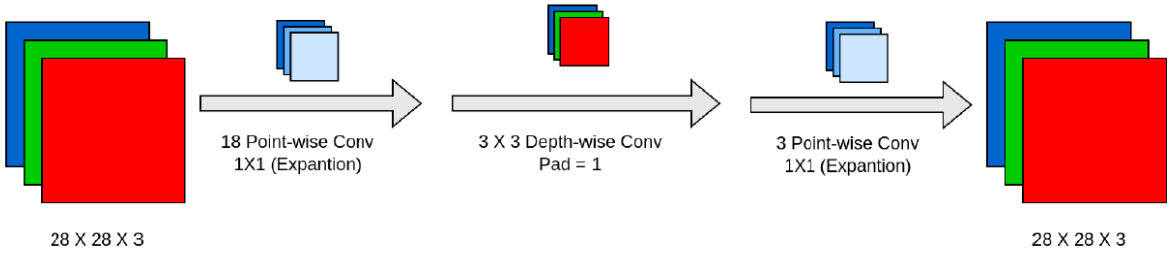
(الف) در شکل ۱ یک پیمانه از شبکه‌ی MobileNet-V1 با یک لایه از شبکه‌ی CNN معمولی نمایش داده شده است. در هر روش تعداد پارامترها را محاسبه کنید.



شکل ۱: لایه‌ی کانولوشن معمولی (پایین) و یک لایه از MobileNet-V1 (بالا).

(ب) در شبکه‌ی MobileNet-V2 از دو فیلتر کانولوشن 1×1 استفاده شده است. شمای کلی معماری استفاده شده از این شبکه در شکل ۲ نشان داده شده است. با محاسبه‌ی تعداد پارامترها برای این شبکه، این شبکه را با شبکه‌ی MobileNet-V1 مقایسه نمایید.

(ج) به نظر شما بین یک شبکه‌ی CNN معمولی، MobileNet-V1 و MobileNet-V2 کدام یک نسبت به حملات مقاوم‌تر است؟



شکل ۲: یک لایه از MobileNet-V2.

سوال ۲ بهینگی FGSM (۵۰ نمره)

(برای حل سوال یک سری راهنمایی در انتهای سوال قرار گرفته‌است. در صورتی که در حل مسئله بدون آن‌ها به مشکل خوردید می‌توانید از آن‌ها کمک بگیرید.)

(الف) می‌دانیم در صورتی که حمله‌ی خصمانه بدون هدف [۱] باشد، برای داده و برچسب (x, y) برای دسته‌بند داده شده‌ی h با تابع هزینه‌ی $l(h(x), y)$ تابع هدفی که برای حمله می‌نویسیم به صورت زیر است:

$$\max_{||\delta|| \leq \epsilon} l\big(h(x), y\big)$$

که $||\delta||$ نرم بینهایت δ است.

در حمله‌ی FGSM برای دستیابی به نقطه‌ی بهینه‌کننده‌ی این تابع، از قاعده‌ی زیر استفاده می‌شود:

$$x' = x + \epsilon sign\Big(\nabla_x l\big(h(x), y\big)\Big)$$

تابع هدف و الگوریتم FGSM را به گونه‌ای تغییر دهید تا حمله هدفمند [۲] به دسته‌ی y' باشد. به نظر شما با در نظر گرفتن یک ϵ ثابت، حمله‌ی هدفمند موفق‌تر است یا حمله‌ی بدون هدف؟

(ب) حال می‌خواهیم ثابت کنیم که در صورتی که مدل پیش‌بینی کننده خطی باشد برای تمامی توابع هزینه‌ی محدب، حمله‌ی FGSM بهترین حمله خواهد بود. اما این بار از خواص توابع محدب و شروط K.T.T استفاده می‌کنیم. در ابتدا تابع هزینه و شروط مسئله را در فرمت K.K.T بنویسید و متغیرهای K.K.T را تشکیل دهید. در تشکیل رابطه، متغیر مربوط به δ_i را α_i در نظر بگیرید.

(ج) از دوگان کمک بگیرید و با قرار دادن مشتق عبارت لاگرانژ نسبت به δ_i برابر با صفر، یک رابطه برای α_i بر حسب متغیرهای اصلی مسئله به دست آورید.

(د) میدانیم که $\delta_i = sign(\delta_i)|\delta_i|$. با قرار دادن α_i به دست آمده در شروط وجود دوگان K.K.T (راهنمایی ۱) رابطه‌ای برای δ_i به دست آورید. (راهنمایی ۲)

(ه) ثابت کنید عبارت به دست آمده با عبارتی که در قسمت (ب) نوشته‌اید معادل است.

(راهنمایی ۱) - شرط چهارم (Complementary slackness). (راهنمایی ۲) - ثابت کنید $sign(\alpha_i) = I(\alpha_i \neq 0)$

[۱] untargeted
[۲] targeted

سوال ۳ چگونگی عمکرد **Universal Adversarial Perturbations** (۵۰ نمره)

(الف) گام‌های حمله‌ی UAP را ذکر کنید و آن‌ها رو توضیح دهید.

(ب) توصیف کنید که چرا روش بالا تنها در ابعاد بالا می‌تواند کار کند؟

(ج) عملکرد تابع را به صورت تجربی در ابعاد پایین بر روی محیط یک دیتاست دوبعدی نظیر two moons بررسی کنید. یعنی در یک ناحیه‌ی کراندار جهات مختلف را در نظر بگیرید و موفقیت حمله در صورت اعمال perturbation به همه‌ی نقاط در آن جهت را به اندازه‌ی epsilon بسنجید و نتیجه‌ی خود را تحلیل کنید. سعی کنید با پارامترهای دیتاست بازی کنید تا یک Failure Scenario در ابعاد پایین بیابید.

سوال ۴ تمرین عملی (۸۰ نمره)

نوت‌بوک adversarial.ipynb را کامل کنید.

موفق باشید.