



دانشگاه صنعتی امیرکبیر
دانشکده مهندسی کامپیوتر
پروژه درس رایانش عصبی و یادگیری عمیق



پروژه هشتم

هدف: آشنایی با حملات و آموزش خصمانه، شبکه‌های مقاوم، معماری‌های رمزگذار و رمزگشا

کد: پیاده سازی این پروژه را به زبان پایتون انجام دهید؛ در این فعالیت مجاز به استفاده از tensorflow یا pytorch یا jax می‌باشید. فایل‌های کد خود را بر اساس شماره سوال و زیر قسمت خواسته شده‌ی آن نام گذاری کنید (برای مثال می‌توان نام گذاری قسمت اول برای سوال سوم تمرین را بصورت `P3_a_preprocessing.py` در نظر گرفت). فایل‌های ارسالی‌تان بایستی با فرمت `py` یا `ipynb` (با حفظ خروجی هر سلول) باشد.

گزارش: ملاک اصلی انجام فعالیت، گزارش آن است و ارسال کد بدون گزارش فاقد ارزش است. برای این فعالیت یک فایل گزارش در قالب pdf تهیه کنید که دارای فهرست بوده و پاسخ‌ها بترتیب در آن قرار گرفته اند و نام، نام خانوادگی و شماره دانشجویی‌تان در قسمت چپ سربرگ تمامی صفحات تکرار شده است. علاوه بر خواسته‌ی مستقیم هر سوال، مقتضی است که نمودارهای خطا (loss) و صحت (accuracy) را به ازای مجموعه داده‌های آموزش و اعتبارسنجی رسم نمایید. همچنین در صورت امکان ماتریس درهم‌ریختگی را بصورت رنگ‌آمیزی شده به همراه اعداد متناظر برای مجموعه داده‌های آموزش، آزمون و اعتبارسنجی نیز تولید نمایید. لازم به ذکر است که در هر آموزش بایستی موارد مهم تنظیم شده نظیر تابع خطا، بهینه‌ساز (به همراه پارامترهای تنظیم شده‌ی آن مانند نرخ یادگیری)، معماری شبکه‌ی آموزشی (کتابخانه‌ها و ابزارهایی برای بصری‌سازی موجود است)، تعداد گام آموزشی، اندازه دسته (Batch Size)، آمارگان تفکیک مجموعه داده (به آموزش، آزمون و اعتبارسنجی)، پیش‌پردازش‌های اعمالی بروی داده‌گان ورودی و... ذکر گردد.

تذکر: مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیر مجاز بوده و با تمامی طرفین برخورد خواهد شد. استفاده از کدها و توضیحات اینترنت به منظور یادگیری صرفاً با ارجاع به آن بلامانع است، اما کپی کردن آن غیرمجاز است.

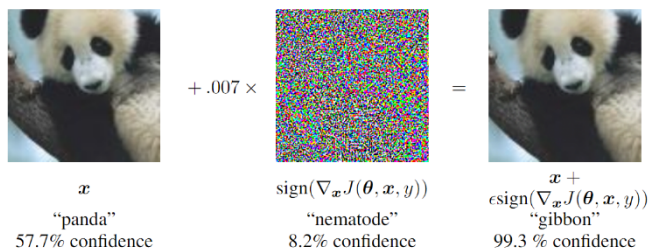
راهنمایی: در صورت نیاز می‌توانید سوالات خود را در خصوص پروژه از تدریس‌یارهای درس، از طریق ایمیل زیر یا در گروه تلگرامی بپرسید. [لینک گروه تلگرامی](#)

Email: ann.ceit.aut@gmail.com CC: m.ebadpour@aut.ac.ir

توجه: می‌توانید از منابع و بسترهای سخت افزاری برخط رایگان نظیر Google Colab یا Kaggle استفاده نمایید.

تاخیر مجاز: در طول ترم، ده روز زمان مجاز تاخیر برای ارسال پروژه‌ها در اختیار دارید (بدون کسر نمره). این تاخیر را می‌توانید بر حسب نیاز بین پروژه‌ها مختلف تقسیم کنید که مجموع آن نباید بیشتر از ده روز شود. پس از استفاده از این تاخیر مجاز، هر روز تاخیر باعث کسر ۱۰٪ نمره‌ی کسب شده‌ی آن تمرین خواهد شد.

ارسال: فایل‌های کد و گزارش خود را در قالب یک فایل فشرده با فرمت StudentID_HW08.zip تا تاریخ ۱۴۰۳/۰۴/۲۶ صرفاً از طریق سایت کورسز ارسال نمایید. ارسال از طریق تلگرام، ایمیل و سایر راه‌های ارتباطی مجاز نبوده و تصحیح صورت نخواهد گرفت.



$$x + 0.007 \times \text{sign}(\nabla_x J(\theta, x, y)) = x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$$

x
 "panda"
 57.7% confidence

$\text{sign}(\nabla_x J(\theta, x, y))$
 "nematode"
 8.2% confidence

$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$
 "gibbon"
 99.3 % confidence

حملات خصمانه^۱ نوعی از حملات بر روی مدل‌های یادگیری ماشین به منظور فریب دادن مدل با استفاده از ورودی‌های دستکاری شده است. هدف اصلی این حملات تغییر خروجی مدل به صورت اشتباه است. به سوالات زیر پاسخ دهید و به منبع یا منابعی که استفاده کردید [ارجاع](#) دهید.

۱- FGSM یکی از اولین و ساده‌ترین روش‌های حمله خصمانه است که توسط یان گودفلو و همکارانش^۲ معرفی شد. هدف این روش، ایجاد یک نمونه خصمانه است که تفاوت بسیار کمی با ورودی اصلی داشته باشد اما مدل را به اشتباه بیندازد. PGD یک روش قوی‌تر و بهبود یافته نسبت به FGSM است که توسط Madry و همکارانش^۳ معرفی شد. این روش به جای انجام یک مرحله، بروزرسانی‌های متعددی را انجام می‌دهد و در هر مرحله، تغییرات را در محدوده‌ای مشخص پروژکت می‌کند تا اطمینان حاصل شود که نمونه خصمانه بیش از حد از ورودی اصلی فاصله نگیرد. این دو روش را مطالعه و خلاصه‌ای از آن‌ها بنویسید. (۱۰ نمره)

۲- چگونه آموزش خصمانه^۴ می‌تواند بر تعمیم‌پذیری مدل به داده‌های دیده نشده تاثیر بگذارد؟ آیا همیشه بهبود در مقاوم شدن در برابر حملات، بهبود صحت بروی داده‌های دیده نشده را تضمین می‌کند؟ نشان دهید. (۱۰ نمره)

۳- چرا و چگونه نمونه‌های خصمانه ایجاد شده برای یک مدل می‌توانند مدل‌های دیگر را نیز فریب دهند؟ این خاصیت انتقال‌پذیری چگونه می‌تواند در حملات جعبه سیاه استفاده شود؟ (۵ نمره)

۴- چگونه می‌توان حملات خصمانه را در حوزه‌ای مانند پردازش زبان طبیعی پیاده‌سازی کرد؟ چه چالش‌های خاصی در این حوزه وجود دارد؟ (۵ نمره)

۵- چگونه می‌توان آموزش خصمانه را در مجموعه داده‌های نامتوازن پیاده‌سازی کرد و چه چالش‌هایی در این مسیر وجود دارد؟ (۵ نمره)

۶- در این سوال می‌خواهیم یک حمله خصمانه با روش‌های FGSM طراحی کنیم و سپس مدل از پیش آموزش داده شده ResNet18 را با آموزش خصمانه مقاوم سازیم. به این منظور مراحل زیر را دنبال کنید (۳۵ نمره):

الف- مدل از پیش آموزش دیده ResNet18 را برای مجموعه داده CIFAR10 آموزش دهید. نمودار خطا آموزش و آزمون را رسم کنید.

¹ Adversarial Attacks

² Explaining and Harnessing Adversarial Examples

³ Towards Deep Learning Models Resistant to Adversarial Attacks

⁴ Adversarial Training

ب- روش FGSM را پیاده‌سازی کنید و ۵ تصویر را به صورت تصادفی انتخاب کنید و به مدل حمله کنید سپس برای این تصاویر: تصویر اصلی، تصویر آشفته شده^۵، برچسب اصلی و برچسب پیش‌بینی شده بر روی تصویر آشفته شده را نمایش دهید.

ج- حال با گنجاندن نمونه‌های خصمانه در فرآیند آموزش، مدل ResNet18 را دوباره آموزش دهید (آموزش خصمانه). این فرآیند به مدل کمک می‌کند تا در برابر حملات خصمانه مقاوم‌تر شود. نحوه آموزش را کامل شرح دهید. نمودارهای زیر را در کنارهم رسم و تفسیر کنید.

- train – natural: خطای آموزش بر روی مدل طبیعی
- train – adversary: خطای آموزش بر روی مدل خصمانه
- test – natural: خطای آموزش بر روی مدل طبیعی (مجموعه داده آزمون بدون تغییر)
- test – adversary: خطای آموزش بر روی مدل خصمانه (مجموعه داده آزمون بدون تغییر)

د- تا اینجا ما توانستیم تا با حملات خصمانه تصویری که تفاوت بسیار کمی با دیتای اصلی دارد، مدل را به اشتباه بیاندازیم. حال می‌خواهیم به صورت هدفمند اینکار را انجام دهیم؛ یعنی مدل باید به اشتباه کلاس مورد نظر ما را پیش‌بینی کند^۶. با روش FGSM حمله هدفمند را پیاده‌سازی و نحوه انجام آن را بطور کامل شرح دهید. حال با ایجاد نمونه‌های خصمانه جدید از مجموعه داده آزمون و همچنین داده‌های آزمون بدون تغییر، صحت هر دو مدل را (مدل طبیعی و مدل آموزش دیده به صورت خصمانه) را ارزیابی کنید. نتایج را تفسیر کنید. در مورد اثربخشی آموزش خصمانه در بهبود استحکام مدل در برابر حملات خصمانه بحث کنید.

۷- در این سوال تصمیم داریم تا برای تصاویر ایرانی یک مدل با معماری رمزگذار و رمزگشا برای وظیفه شرح تصویر^۷ طراحی کنیم. مجموعه داده persian_image_captioning.rar در اختیار شما قرار گرفته است. این مجموعه داده حدود ۱۵۰۰ مقاله خبری به همراه تصاویر مرتبط آن است. این مقالات از سایت خبرگزاری تسنیم جمع‌آوری شده است. فایل news.json حاوی لیستی از اشیاء json که هر کدام دارای اطلاعات زیر هستند:

- title: عنوان مقاله خبری.
- Description: شرح کوتاهی از مقاله.
- Category: دسته‌ای که مقاله به آن تعلق دارد.
- Reporter: نام خبرنگاری که این مطلب را منتشر کرده است.
- Time: تاریخ و ساعتی که مقاله در آن منتشر شده است.
- Images: لیستی از تصاویر مرتبط با مقاله (همه آنها را می‌توانید در پوشه images پیدا کنید).

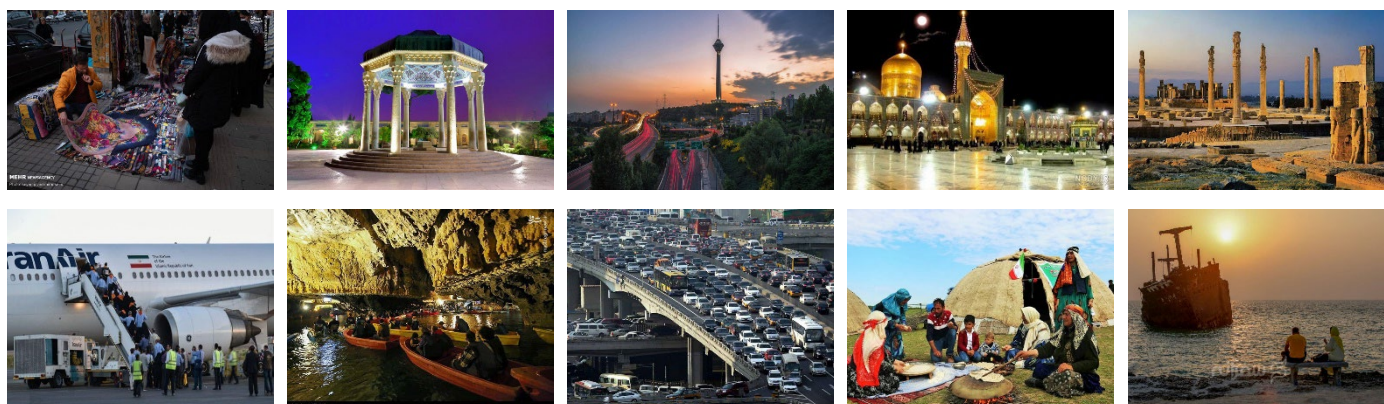
⁵ Perturbed

⁶ Targeted Attacks

⁷ Image Captioning

عنوان هر مقاله را می‌توان به عنوان یک شرح (caption) برای تصاویر مرتبط با آن مقاله، در نظر گرفت. همچنین می‌توانید با جایگزین کردن مترادف کلمات و همچنین با روش‌های دلخواه برای تصاویر، داده افزایی^۸ کنید. در نهایت مدلی آموزش دهید تا این وظیفه را انجام دهد. موارد زیر را در گزارش خود لحاظ و توضیح کامل دهید:

- پیش‌پردازی که انجام دادید.
- معماری مدل پیشنهادی خود را رسم کنید.
- تابع هزینه‌ای^۹ که استفاده کردید.
- روش‌هایی که برای ارزیابی این وظیفه استفاده کردید.
- اسکریپتی بنویسید تا با دریافت مسیر یک پوشه، شرح تصاویر در آن پوشه را در یک فایل txt بنویسد. پوشه تحت عنوان `selected_images` در اختیار شما قرار گرفته است. مسیر این پوشه را به اسکریپت خود بدهید و خروجی آن را (شرح تصاویر) همراه با تصاویر مرتبط ارسال کنید. دقت کنید که اسکریپت نوشته شده توسط شما در روز تحویل پروژه توسط تصاویر دیگر بررسی خواهد شد. تصاویر این پوشه در زیر نشان داده شده است:



توجه فرمایید نمره این تمرین (۳۰ + ۳۰ امتیازی) است. یعنی در صورتی که مراحل پیش‌پردازش، معماری مدل، صحت‌نهایی و به طور کلی روش حل مسئله، دارای خلاقیت و کیفیت مورد قبولی باشد، علاوه بر نمره اصلی تا ۳۰ امتیاز، نمره اضافی برای شما در نظر گرفته خواهد شد.

موفق باشید

⁸ Data Augmentation

⁹ Loss Function