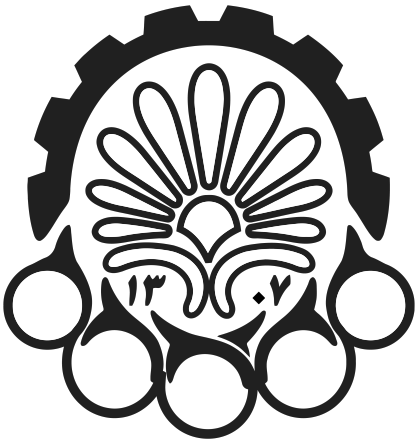
شبکه‌های عصبی و یادگیری عمیق
دکتر صفابخش



دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

رضا آدینه پور ۴۰۲۱۳۱۰۵۵

تمرین هشتم

ساختارهای Encoder و Decoder

۲۲ تیر ۱۴۰۳

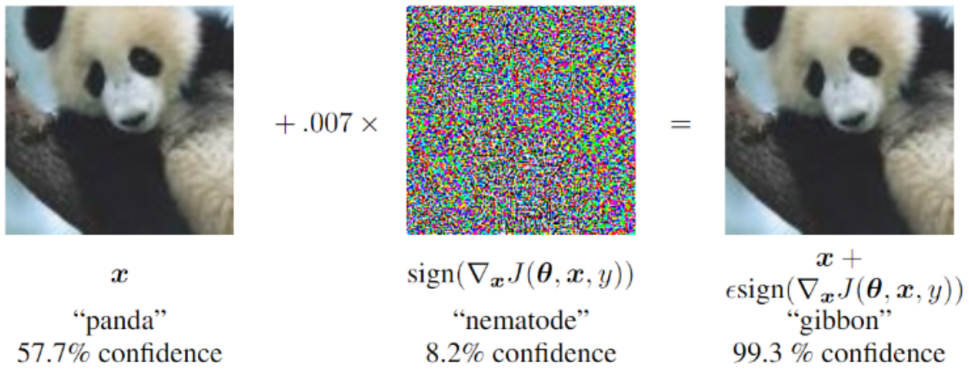
شبکه‌های عصبی و یادگیری عمیق

تمرین هشتم

رضا آدینه پور ۴۰۲۱۳۱۰۵۵

دانشکده مهندسی کامپیوتر

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.



شکل ۱: تغییر نمونه ورودی

سوال اول - تئوری

یکی از اولین و ساده ترین روش‌های حمله خصمانه، FGSM 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 شده. این روش به جای انجام یک مرحله، بروز رسانی‌های متعددی را انجام می‌دهد و در هر مرحله تغییرات را در محدوده مشخصی پروجکت 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.

پاسخ

[۱] Adversarial Attack

[۲] ExamplesAdversarial Harnessing and Explaining

[۳] AttacksAdversarial to Resistant Models Learning Deep Towards

سوال دوم - تئوری

چگونه آموزش خصمانه[۴] می‌تواند بر تعمیم پذیری مدل به داده‌های دیده نشده تاثیر بگذارد؟ آیا همیشه بهبود در مقاومت شدن در برابر حملات، بهبود صحت بر روی داده‌های دیده نشده را تضمین می‌کند؟ نشان دهید.

پاسخ

[۴] Adversarial Training

سوال سوم - تئوری

چرا و چگونه نمونه‌های خصمانه‌ی ایجاد شده برای یک مدل می‌توانند مدل‌های دیگر را نیز فریب دهند؟ این خاصیت انتقال‌پذیری چگونه می‌تواند در حملات جعبه سیاه استفاده شود؟

پاسخ

سوال چهارم - تئوری

۴- چگونه می‌توان حملات خصمانه را در حوزه‌هایی مانند پردازش زبان طبیعی پیاده‌سازی کرد؟ چه چالش‌های خاصی در این حوزه وجود دارد؟

پاسخ

سوال پنجم - تئوری

چگونه می‌توان آموزش خصمانه را در مجموعه داده‌های نامتوازن پیاده‌سازی کرد و چه چالش‌هایی در این مسیر وجود دارد؟

پاسخ

سوال ششم - تئوری

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:

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

۲. 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]، پرچسب اصلی و پرچسب پیش‌بینی شده بر روی تصویر آشفته شده را نمایش دهید.

۳. 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 کنار هم رسم و تفسیر کنید.

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

۴. تا اینجا ما توانستیم تا با حملات خصمانه تصویری که تفاوت بسیار کمی با دیتای اصلی دارد، مدل را به اشتباه بیندازیم. 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.

پاسخ

[۵] Perturbed
[۶] Target Attack