

# 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

۲۳ تیر ۱۴۰۳



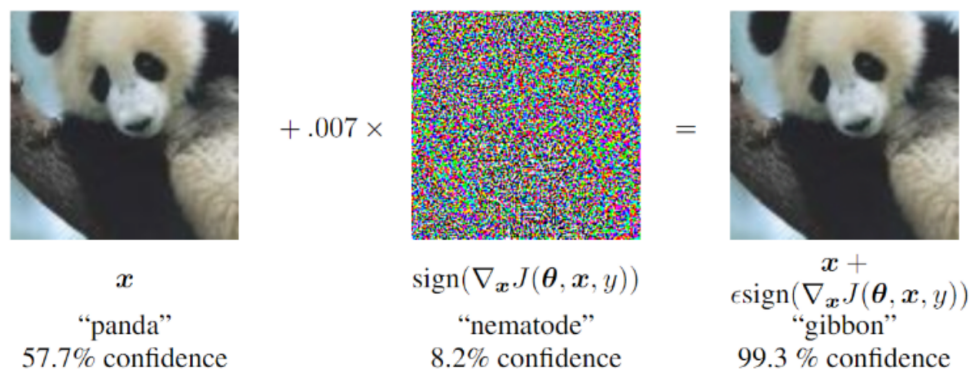
دانشکده مهندسی کامپیوتر

# شبکه‌های عصبی و یادگیری عمیق

تمرین هشتم

رضا آدینه پور ۴۰۲۱۳۱۰۵۵

حملات خصمانه<sup>۱</sup> 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 ساده‌ترین روش‌های حمله خصمانه، FGSM است که توسط یان گودفلو و همکارانش<sup>۲</sup> 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<sup>۳</sup> 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

### • روش FGSM:

این روش در سال ۲۰۱۵ در مقاله Explaining and Harnessing Adversarial Examples معرفی شد که یکی از ساده‌ترین روش‌های حملات خصمانه است. همانطور که در صورت سوال نیز توضیح داده شد، هدف اصلی FGSM ایجاد نمونه‌های خصمانه‌ای است که از نظر بصری تفاوت زیادی با ورودی اصلی نداشته باشند ولی باعث شوند مدل یادگیری ماشین خطا کند.

<sup>۱</sup> Adversarial Attack

<sup>۲</sup> Examples Adversarial Harnessing and Explaining

<sup>۳</sup> Attacks Adversarial to Resistant Models Learning Deep Towards

## پاسخ

در ادامه، روش انجام این الگوریتم را به صورت خلاصه توضیح می‌دهیم:

- FGSM با اضافه کردن یک اختلال به داده‌های ورودی اصلی کار می‌کند.
- اختلال با استفاده از گرادیان تابع هزینه نسبت به داده‌های ورودی محاسبه می‌شود.
- به طور خاص، اختلال به صورت زیر محاسبه می‌شود:

$$\eta = \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$$

که در آن:

- $\epsilon$  یک مقدار عددی کوچک است که میزان اختلال را کنترل می‌کند.
- $\nabla_x J(\theta, x, y)$  گرادیان تابع هزینه  $J$  نسبت به ورودی  $x$  است.
- $\text{sign}$  تابع علامت است که علامت گرادیان را استخراج می‌کند.
- سپس نمونه‌ی خصمانه با اضافه کردن این اختلال به ورودی اصلی ایجاد می‌شود:

$$x' = x + \eta$$

## سوال دوم - تئوری

چگونه آموزش خصمانه<sup>۴</sup> می‌تواند بر تعمیم‌پذیری 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

<sup>۴</sup>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 - عملی

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<sup>۵</sup>، 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 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)

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<sup>۶</sup>. 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

<sup>۵</sup> Perturbed  
<sup>۶</sup> Target Attack



## سوال هفتم - عملی

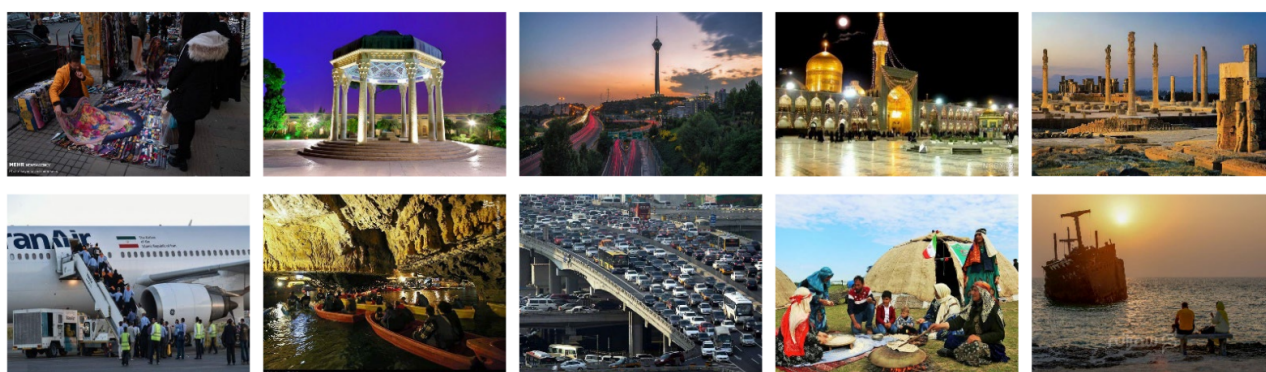
در این سوال تصمیم داریم تا برای تصاویر ایرانی یک مدل با معماری رمزگذار و رمزگشا برای وظیفه شرح تصویر<sup>۷</sup> طراحی کنیم. مجموعه داده persian\_image\_captioning.rar در اختیار شما قرار گرفته است. این مجموعه داده حدود ۱۵۰۰ مقاله خبری به همراه تصاویر مرتبط آن است. این مقالات از سایت خبرگزاری تسنیم جمع‌آوری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

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

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<sup>۸</sup> کنید. در نهایت مدلی آموزش دهید تا این وظیفه را انجام دهد. موارد زیر را در گزارش خود لحاظ و توضیح کامل دهید:

۱. پیش‌پردازشی که انجام داده‌اید.
۲. معماری مدل پیشنهادی خود را رسم کنید.
۳. تابع هزینه‌ای<sup>۹</sup> که استفاده کردید.
۴. روش‌هایی که برای ارزیابی این وظیفه استفاده شده.

اسکرپیتی 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 اسکرپیت خود بدهید و خروجی آن را (شرح تصاویر) همراه با تصاویر مرتبط ارسال کنید. دقت کنید که اسکرپیت نوشته شده توسط شما در روز تحویل پروژه توسط دیگر تصاویر بررسی خواهد شد. تصاویر این پوشه در زیر نشان داده شده است:



شکل ۲: تصاویر پوشه selected\_images

Image Captioning<sup>۷</sup>  
Data Augmentation<sup>۸</sup>  
Function Loss<sup>۹</sup>

توجه فرمایید نمره این تمرین  $(30 + 30)$  امتیازی است. یعنی در صورتی که مراحل پیش‌پردازش، معماری مدل، صحت نهایی 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.

پاسخ