

# شبکه‌های عصبی و یادگیری عمیق

## دکتر صفا بخش



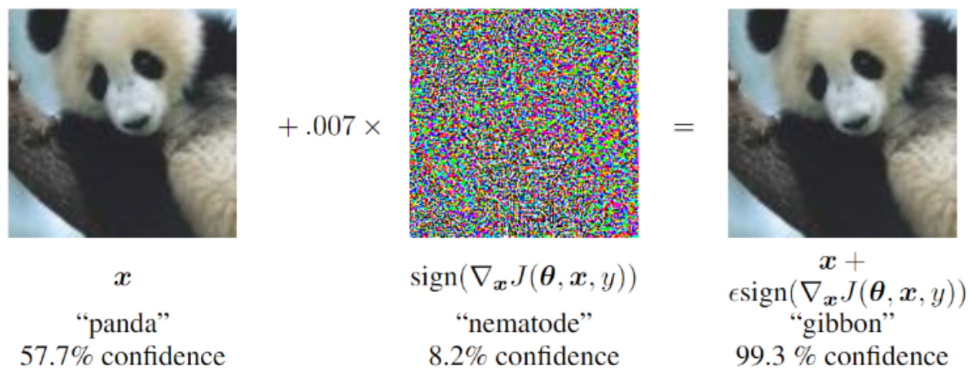
**دانشگاه صنعتی امیرکبیر**  
( پلی تکنیک تهران )  
دانشکده مهندسی کامپیوتر

رضا آدینه پور ۴۰۲۱۳۱۰۵۵

تمرین هشتم  
ساختارهای Encoder و Decoder

۲۵ تیر ۱۴۰۳

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.
- این اختلال با استفاده از گرادیان تابع هزینه نسبت به داده‌های ورودی محاسبه می‌شود.
- به طور خاص، می‌توان فرمول این اختلال را به صورت زیر بیان نمود:

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

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:

$$x' = x + \eta$$

از مزایا و معایب FGSM می‌توان به موارد زیر اشاره کرد:  
مزایا: از نظر محاسباتی کارآمد و ساده است و به همین دلیل انتخاب محبوبی برای مطالعات اولیه در مورد حملات خصمانه است.  
معایب: سادگی FGSM می‌تواند به عنوان یک نقطه ضعف نیز عمل کند، زیرا اختلالات ایجاد شده ممکن است در مقابل مدل‌های مقاوم‌تر به اندازه‌ی کافی مؤثر نباشند.

### • روش PGD:

همانطور که در صورت سوال نیز گفته شد، روش PGD یک روش بهبود یافته نسبت به FGSM است که در سال ۲۰۱۷ در مقاله‌ی *Towards Deep Learning Models Resistant to Adversarial Attacks* معرفی شد. هدف اصلی این روش، ایجاد نمونه‌های خصمانه قوی‌تر با انجام چندین به‌روزرسانی گرادیانی و اطمینان از ماندن اختلالات در یک محدوده‌ی مشخص است.  
در ادامه، روش انجام این الگوریتم را توضیح می‌دهیم:

- PGD به طور مکرر نمونه‌ی خصمانه را با گرفتن چندین گام گرادیانی اصلاح می‌کند.
- هر گام شامل مراحل زیر است:

۱. محاسبه‌ی گرادیان: گرادیان تابع هزینه نسبت به نمونه‌ی خصمانه‌ی فعلی محاسبه می‌شود.
۲. گام به‌روزرسانی: نمونه‌ی خصمانه با حرکت در جهت گرادیان به‌روزرسانی می‌شود:

$$x_{t+1} = x_t + \alpha \cdot \text{sign}(\nabla_x J(\theta, x_t, y))$$

که  $\alpha$  اندازه‌ی گام است.

۳. پروژه‌سازی: اطمینان حاصل می‌شود که نمونه‌ی خصمانه‌ی به‌روز شده در یک کره‌ی  $\epsilon$  حول ورودی اصلی باقی می‌ماند:

$$x_{t+1} = \text{clip}(x_{t+1}, x - \epsilon, x + \epsilon)$$

این مرحله اطمینان می‌دهد که اختلال از محدوده‌ی مجاز تجاوز نمی‌کند.

پاسخ

این فرآیند برای تعداد معینی از تکرارها یا تا همگرایی تکرار می‌شود.  
از مزایا و معایب PGD می‌توان به موارد زیر اشاره کرد:

۱. مزایا: PGD به دلیل اینکه فضای اختلالات ممکن را با تکرارهای متعدد به طور جامع‌تری کاوش می‌کند، به عنوان یک روش حمله‌ی قوی‌تر نسبت به FGSM شناخته می‌شود.
۲. معایب: طبیعت تکراری PGD باعث می‌شود که از نظر محاسباتی پرهزینه‌تر از FGSM باشد.

\*  
—————

## References

- [1] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In Proceedings of the International Conference on Learning Representations (ICLR) [\[Link\]](#)
- [2] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In Proceedings of the International Conference on Learning Representations (ICLR) [\[Link\]](#)

## 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

## ۱. توازن بین مقاومت و دقت:

- بهبود مقاومت: آموزش خصمانه به طور عمده مقاومت مدل در برابر حملات خصمانه را بهبود می‌بخشد. با معرفی نمونه‌های خصمانه به مدل در طول آموزش، مدل یاد می‌گیرد که حتی در مواجهه با ورودی‌های دستکاری شده نیز پیش‌بینی‌های دقیقی انجام دهد.
- دقت در داده‌های تمیز: در حالی که آموزش خصمانه می‌تواند مقاومت را به طور قابل توجهی افزایش دهد، اغلب با کاهش دقت در داده‌های تمیز (غیر دستکاری شده) همراه است. این به این دلیل است که مدل ممکن است بیش از حد بر روی نمونه‌های خصمانه تمرکز کند و توانایی تعمیم‌دهی به نمونه‌های عادی و غیر خصمانه کاهش یابد.

## ۲. تعمیم‌پذیری به داده‌های دیده نشده:

- توانایی تعمیم‌دهی: در برخی موارد، مقاومتی که توسط آموزش خصمانه ایجاد می‌شود، می‌تواند به مدل کمک کند تا به داده‌های دیده نشده خاصی بهتر تعمیم دهد، به ویژه اگر داده‌های دیده نشده شامل نویز یا تغییرات مشابه نمونه‌های خصمانه باشد. با این حال، اگر داده‌های دیده نشده تمیز باشند، ممکن است مدل نسبت به مدلی که بدون نمونه‌های خصمانه آموزش دیده، عملکرد ضعیف‌تری داشته باشد.
- تنوع توزیع داده‌ها: قابلیت تعمیم‌دهی مدل همچنین به این بستگی دارد که نمونه‌های خصمانه چقدر نماینده تغییرات احتمالی در داده‌های دیده نشده هستند. اگر نمونه‌های خصمانه به خوبی طراحی شده باشند و طیف وسیعی از تغییرات ممکن را پوشش دهند، می‌توانند به مدل کمک کنند تا بهتر تعمیم دهد. در مقابل، اگر نمونه‌های خصمانه بیش از حد خاص باشند، ممکن است مدل با توزیع داده‌های جدیدی که به خوبی نمایندگی نشده‌اند، مشکل داشته باشد.

بهبود مقاومت در برابر حملات خصمانه در مقابل دقت بر روی داده‌های دیده نشده همیشه تضمین نمی‌شود، بهبود مقاومت در برابر حملات خصمانه همیشه دقت بهتر بر روی داده‌های دیده نشده را تضمین نمی‌کند. فرآیند آموزش خصمانه می‌تواند گاهی منجر به overfitting به نمونه‌های خصمانه شود که ممکن است به خوبی به نمونه‌های جدید و دیده نشده تعمیم ندهد. دستیابی به تعادل بین مقاومت و دقت نیاز به توجه دقیق به داده‌های آموزشی و انواع نمونه‌های خصمانه مورد استفاده دارد. تکنیک‌های regularization، data augmentation و سایر روش‌ها می‌توانند به کاهش تأثیر منفی بر دقت کمک کنند در حالی که مقاومت را نیز بهبود می‌بخشند.

<sup>۴</sup>Adversarial Training

## سوال سوم - تئوری

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

پاسخ

یک جنبه جالب و مهم از نمونه‌های خصمانه خاصیت انتقال‌پذیری آنهاست، یک نمونه خصمانه که برای فریب یک مدل طراحی شده است، می‌تواند مدل‌های دیگر را نیز فریب دهد. از این پدیده در حملات جعبه سیاه استفاده می‌شود، جایی که مهاجم دسترسی مستقیم به مدل هدف ندارد. این موضوع می‌تواند به دلایل زیر باشد:

۱. شباهت در ویژگی‌های آموخته شده

۲. داده‌های آموزشی مشترک

۳. آسیب‌پذیری‌های مشترک:

مدل‌های مختلف، به ویژه آنهایی که بر روی همان مجموعه داده یا با معماری‌های مشابه آموزش دیده‌اند، تمایل به یادگیری ویژگی‌های مشابه دارند. مثال‌های خصمانه از ضعف‌های این ویژگی‌های آموخته شده سوءاستفاده می‌کنند، که باعث می‌شود به احتمال زیاد بین مدل‌ها انتقال یابند.

۴. ماهیت خطی مدل‌ها:

یساری از مدل‌های یادگیری ماشین، به ویژه شبکه‌های عصبی، طبیعتی خطی دارند. تغییرات خصمانه از این خطی بودن سوءاستفاده می‌کنند و از آنجا که این ویژگی بین مدل‌ها مشترک است، مثال‌های خصمانه می‌توانند انتقال پیدا کنند.

۵. همپوشانی در مرزهای تصمیم‌گیری:

مدل‌هایی که بر روی داده‌های مشابه آموزش دیده‌اند، تمایل به داشتن مرزهای تصمیم‌گیری همپوشان دارند. مثال‌های خصمانه‌ای که برای عبور از مرز تصمیم‌گیری یک مدل ساخته شده‌اند، به احتمال زیاد از مرزهای تصمیم‌گیری مدل‌های دیگر که بر روی داده‌های مشابه آموزش دیده‌اند نیز عبور می‌کنند.

در حملات جعبه سیاه، مهاجم دسترسی مستقیم به پارامترهای مدل هدف یا داده‌های آموزشی آن ندارد. در عوض، آنها از خاصیت انتقال‌پذیری مثال‌های خصمانه بهره می‌برند. به همین دلیل می‌توان مدلی جایگزین را آموزش داد که مدل هدف را تقریب بزند. این کار می‌تواند با استفاده از داده‌های مشابه یا تقلید از رفتار مدل هدف انجام شود.

هنگامی که مدل جایگزین آموزش دید، نمونه‌های خصمانه با استفاده از این مدل ساخته می‌شوند. سپس این نمونه‌های ساخته شده، به مدل هدف اعمال می‌شوند. به دلیل خاصیت انتقال‌پذیری، این نمونه‌ها احتمال بالایی برای فریب مدل هدف دارند.

## 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 گنجانیدن 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

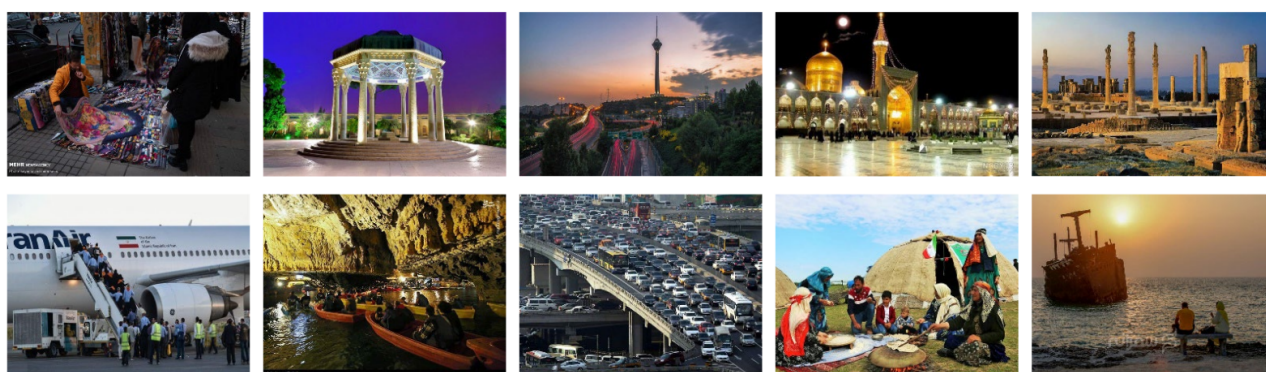
در این سوال تصمیم داریم تا برای تصاویر ایرانی یک مدل با معماری رمزگذار و رمزگشا برای وظیفه شرح تصویر<sup>۷</sup> طراحی کنیم. مجموعه داده persian\_image\_captioning.rar در اختیار شما قرار گرفته است. این مجموعه داده حدود ۱۵۰۰ مقاله خبری به همراه تصاویر مرتبط آن است. این مقالات از سایت خبرگزاری تسنیم جمع‌آوری 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  $(۳۰ + ۳۰)$  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