به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس: هوش مصنوعی قابل اعتماد

مدرس: دکتر مصطفی توسلی پور

تمرین شماره ۱

فروردین ماه ۱۴۰۵

فهرست

سوال اول: تعمیم پذیری ۳

مقدمه ۳

بخش اول: سوالات توضیحی ۴

زیربخش اول: مفهوم همبستگی کاذب و شکستERM (۵ نمره) ٤

زیربخش دوم: فرمول‌سازی خطای بدترین گروه و گروه‌های داده (۵ نمره) ٤

زیربخش سوم: نقش تابع زیانGCE (۵ نمره) ٤

زیربخش چهارم: بررسی فرض اول مقاله (۵ نمره) ٥

زیربخش پنجم: ایده الگوریتمJTT (۵ نمره) ٥

بخش دوم: سوالات پیاده سازی ۶

زیربخش اول: آماده سازی داده و مصورسازی (۶ نمره) ٦

زیربخش دوم: آموزش مدل پایه باERM (۴ نمره) ٦

زیربخش سوم: استخراج مجموعه خطا(Error Set) (۶ نمره) ٧

زیربخش چهارم: ساخت دیتاست وزن‌دار(Upsampling) (۶ نمره) ٧

زیربخش پنجم: آموزش مجدد مدل (۸ نمره) ٧

زیربخش ششم: پیاده‌سازی تابع زیانGCE (۵ نمره) ٧

سوال دوم: حملات خصمانه ۹

مقدمه ۹

بخش اول: مرور مفاهیم (۱۰ نمره) ۱۰

بخش دوم: حمله جعبه سیاه (۲۱ نمره) ۱۱

بخش سوم: پیاده سازی حملات جعبه سفید ۱۲

زیربخش اول: مدل و مجموعه داده ۱۲

زیربخش دوم: FGSM (۱۲ نمره) ۱۲

زیربخش سوم: PGD (۱۲ نمره) ۱۳

نکات تحویل ۱۴

سوال اول: تعمیم پذیری

مقدمه

مفهوم تعمیم‌پذیری[1] یکی از ارکان اصلی یادگیری ماشین است که به توانایی مدل در حفظ عملکرد مناسب بر روی داده‌های دیده نشده اشاره دارد. در بسیاری از مسائل دنیای واقعی، مدل‌ها به جای یادگیری الگوهای اصیل و علی[2]، به میانبرها و همبستگی‌های کاذب[3] موجود در داده‌های آموزشی وابسته می‌شوند؛ این امر باعث می‌شود که با وجود خطای نزدیک به صفر در مرحله آموزش، مدل در محیط تست با افت شدید عملکرد مواجه شود. در این سوال قصد داریم ضمن بررسی تئوریک دلایل شکست روش‌های سنتی مانند کمینه سازی ریسک تجربی[4] در مواجهه با همبستگی‌های کاذب، با مطالعه مقالات معتبر این حوزه، رویکردهای کارآمدی را برای استخراج نمونه‌های خطادار و آموزش مدل‌های تعمیم پذیرتر بررسی و پیاده‌سازی کنیم.

[1] Generalization
[2] Causal
[3] Spurious Correlations
[4] ERM

بخش اول: سوالات توضیحی

در این بخش باید مقاله Mitigating Spurious Correlations را مطالعه کرده و به سوالات زیر پاسخ دهید. پاسخ‌های شما باید دقیق، مفهومی و مبتنی بر استدلال‌های مطرح شده در مقاله باشد.

زیربخش اول: مفهوم همبستگی کاذب و شکست ERM (۵ نمره)

با توجه به مقاله، مفهوم همبستگی کاذب را توضیح دهید. چرا روش سنتی بهینه‌سازی یعنی کمینه سازی ریسک تجربی (ERM) در مواجهه با این پدیده شکست می‌خورد و مدل نهایی به جای یادگیری ویژگی‌های اصلی، روی ویژگی‌های کاذب Overfit می‌شود؟

زیربخش دوم: فرموله سازی خطای بدترین گروه و گروه‌های داده (۵ نمره)

در مقاله برای جلوگیری از سوگیری مدل، مسئله به صورت بهینه‌سازی در بدترین شرایط(Worst-Group Loss) فرموله شده است. فرمول min—max ارائه‌شده در مقاله برای این تابع هدف را بنویسید. سپس با توجه به تعاریف مقاله، تفاوت بین نمونه‌های هم‌راستا با بایاس (b_a) و نمونه‌های در تضاد با بایاس (b_c) را به صورت خلاصه شرح دهید.

زیربخش سوم: نقش تابع زیان GCE (۵ نمره)

در الگوریتم پیشنهادی مقاله (DPR)، برای تقریب احتمال عضویت نمونه‌ها در گروه اقلیت (b_c) بدون داشتن برچسب صریح آن‌ها، یک مدل دارای سوگیری[۵] آموزش داده می‌شود. چرا نویسندگان برای آموزش این مدل، به جای تابع زیان استاندارد Cross-Entropy ، از تابع Generalized Cross-Entropy استفاده کرده‌اند؟ رابطه این تابع زیان را هم بنویسید.

---

[۵] Biased Model

زیربخش چهارم: بررسی فرض اول مقاله (۵ نمره)

یکی از پایه‌های اصلی اثبات‌ها و رویکرد مقاله، فرض شماره یک (Assumption 1) است. این فرض چه مفهومی دارد و چرا برای بازنویسی تابع هدف و شناسایی نمونه‌های سخت ضروری است؟

زیربخش پنجم: ایده الگوریتم JTT (۵ نمره)

مقاله (JTT) Just Train Twice ایده ساده‌تری برای شناسایی نمونه‌های b_c و مقابله با همبستگی کاذب ارائه داده است. با مطالعه چکیده و مقدمه مقاله JTT ، در ۳ الی ۵ خط توضیح دهید که ایده اصلی این روش چیست و چگونه در دو مرحله آموزش مدل نهایی را در برابر همبستگی‌های کاذب مقاوم می‌کند؟

بخش دوم: سوالات پیاده سازی

در این بخش قصد داریم یک شبکه عصبی (MLP) را برای دسته‌بندی دیتاست Colored MNIST پیاده‌سازی کنیم. فایل‌های خام و استاندارد دیتاست MNIST (شامل ۴ فایل فشرده) در اختیار شما قرار گرفته است. هدف این است که ابتدا این داده‌های خام را به یک تسک باینری دارای همبستگی کاذب تبدیل کرده و سپس با استفاده از الگوریتم JTT بر این مشکل غلبه کنیم.

زیربخش اول: آماده سازی داده و مصورسازی (۶ نمره)

چهار فایل خام MNIST را بارگذاری کنید. سپس تغییرات زیر را روی داده‌ها اعمال کنید تا دیتاست Colored MNIST ساخته شود:

۱- **تبدیل به تسک باینری:** برچسب اعداد ۰ تا ۴ را به کلاس ۰ و اعداد ۵ تا ۹ را به کلاس ۱ تغییر دهید.

۲-**تزریق همبستگی کاذب (رنگ):** تصاویر تک‌کاناله را به تصاویر سه کاناله (RGB) تبدیل کنید. در مجموعه آموزش (Train) با احتمال ۹۵ درصد، تصاویر کلاس ۰ را به رنگ قرمز و تصاویر کلاس ۱ را به رنگ سبز در آورید. در مجموعه آزمون (Test) با احتمال ۱۰۰ درصد، این همبستگی را کاملاً برعکس کنید (کلاس ۰ سبز و کلاس ۱ قرمز).

پس از ساخت دیتاست نهایی، حداقل ۱۰ تصویر تصادفی از مجموعه آموزش و ۱۰ تصویر تصادفی از مجموعه آزمون را در قالب یک گرید رسم کنید و برچسب (Label) هر تصویر را بالای آن نمایش دهید تا صحت توزیع رنگ‌ها مشخص شود.

زیربخش دوم: آموزش مدل پایه با ERM (۴ نمره)

مدل MLP را با معماری تعیین‌شده تعریف کنید. مدل را با رویکرد استاندارد ERM روی داده‌های آموزشی، آموزش دهید. دقت نهایی را روی مجموعه آموزش و مجموعه آزمون گزارش کنید. اختلاف مشاهده شده بین این دو دقت را چگونه تحلیل می‌کنید؟

زیربخش سوم: استخراج مجموعه خطا (ERROR SET) (۶ نمره)

مدل ERM آموزش‌دیده را مجدداً روی کل داده‌های مجموعه آموزش ارزیابی کنید. اندیس نمونه‌هایی که مدل در پیش‌بینی کلاس آن‌ها دچار خطا شده است را شناسایی کرده و تحت عنوان Error Set ذخیره کنید. تعداد دقیق نمونه‌های موجود در این مجموعه خطا چقدر است؟

زیربخش چهارم: ساخت دیتاست وزن‌دار (UPSAMPLING) (۶ نمره)

یک دیتاست آموزشی جدید بسازید که شامل تمامی داده‌های آموزشی مرحله اول باشد، با این تفاوت که نمونه‌های موجود در Error Set را با ضریب U=50 تکرار (Upsample) کنید. تعداد نمونه‌های دیتاست جدید را چاپ کنید.

زیربخش پنجم: آموزش مجدد مدل (۸ نمره)

یک مدل MLP کاملاً جدید (با وزن‌های مقداردهی اولیه شده و معماری یکسان) بسازید. این مدل را روی دیتاست وزن‌دار (ساخته شده در مرحله قبل) آموزش دهید. دقت این مدل جدید را روی مجموعه آزمون ارزیابی کرده و با نتایج مدل ERM مقایسه کنید. آیا رویکرد JTT در رفع مشکل همبستگی کاذب موفق بوده است یا خیر؟ تحلیل کنید.

زیربخش ششم: پیاده‌سازی تابع زیان GCE (۵ نمره)

تابع زیان Generalized Cross-Entropy (GCE) را پیاده‌سازی کنید. مدل ERM (زیربخش دوم) را این بار با تابع زیان GCE آموزش دهید و بررسی کنید که آیا استفاده از این تابع تاثیری در تغییر اندازه یا کیفیت Error Set استخراج شده دارد یا خیر. نتایج را به صورت مختصر تحلیل کنید.

توجه: برای آنکه نتایج قابل مقایسه باشد، در تمامی مراحل از معماری ثابت زیر برای مدل MLP استفاده کنید:

Layer	Input →Output Dimensions	Activation Function
Input	Color Image: $3 \times 28 \times 28$	-
Flatten	2352→2352	-
Hidden Layer	2352→128	ReLU
Output Layer	128→2	None

توجه: از Optimizer استاندارد (مانند Adam) و Learning Rate مناسب به انتخاب خودتان استفاده کنید.

تنظیمات آموزش:

- تعداد Epoch برای ERM : ۲ دور
- ضریب تکرار (Upsampling) برای JTT: U=50
- تعداد Epoch برای JTT: ۳ دور

سوال دوم: حملات خصمانه

مقدمه

یکی از چالش‌های مهم در یادگیری ماشین، مقاوت[۶] مدل‌ها است. در بسیاری از کاربردهای حساس، کافی است تغییرات بسیار کوچک و نامحسوس در ورودی ایجاد شود تا مدل رفتاری کاملاً اشتباه نشان دهد. این پدیده که با ایجاد تغییرات عمدی و هدفمند در داده‌ها باعث گمراه کردن مدل می‌شود، حمله متخاصم[۷] نام دارد.

حملات متخاصم نشان می‌دهند که مدل‌های یادگیری عمیق حتی با وجود دقت بالا، می‌توانند نسبت به اختلال‌های بسیار کوچک، آسیب‌پذیر باشند. مطالعه و تحلیل این آسیب‌پذیری‌ها نقش مهمی در طراحی مدل‌های قابل‌اعتمادتر، توسعه سامانه‌های مقاوم و درک عمیق‌تر از رفتار شبکه‌های عصبی دارد.

در این تمرین قصد داریم ابتدا مروری بر مفاهیم این حوزه داشته باشیم. پس از آن با یک رویکرد حمله جعبه سیاه[۸] آشنا شویم. در انتها دو روش حمله جعبه سفید[۹] را پیاده‌سازی خواهیم کرد.

---

[۶] Robustness
[۷] Adversarial Attack
[۸] black-box attack
[۹] white-box attack

بخش اول: مرور مفاهیم (۱۰ نمره)

- روش‌های حمله متخاصم معمولا به دو دسته جعبه‌سفید و جعبه‌سیاه دسته‌بندی می‌شوند. این دو رویکرد را توضیح دهید. (۳ نمره)
- چرا مطالعه و تحقیق درباره حملات متخاصم در حالت جعبه‌سیاه حائز اهمیت است؟(حداکثر در ۴ خط توضیح دهید.) (۴ نمره)
- حملات متخاصم معمولا به دو دسته هدف‌دار[۱۰] و غیرهدف‌دار[۱۱] تقسیم می‌شوند. تفاوت اصلی بین این دو دسته را بیان کنید. (۳ نمره)

[۱۰] Targeted

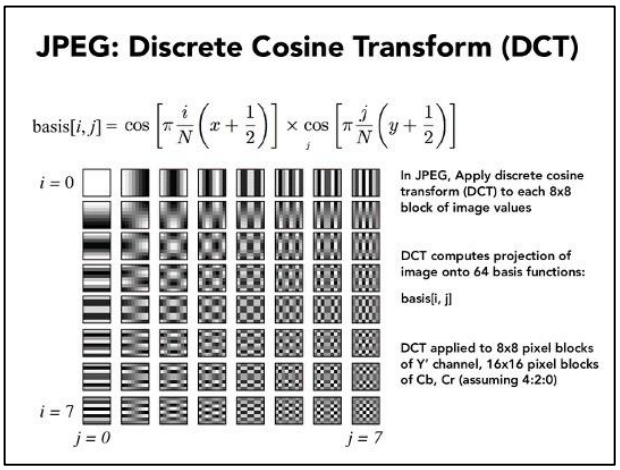
[۱۱] Untargeted

بخش دوم: حمله جعبه سیاه (۲۱ نمره)

در این بخش قصد داریم یک مقاله مربوط به حمله جعبه سیاه را مورد بررسی قرار دهیم. ابتدا مقاله Simple Black-box Adversarial Attacks را مطالعه کرده، سپس به سوال‌های زیر پاسخ دهید.

- با توجه به بخش Introduction مقاله، چرا به نظر می‌رسد همه مدل‌های طبقه‌بندی تصاویر به حملات جعبه‌سفید آسیب‌پذیر هستند؟ مقاله چه استدلالی برای این موضوع بیان می‌کند؟ (حداکثر در ۳ خط توضیح دهید.) (۴ نمره)
- با توجه به بخش Introduction مقاله، دلیل اتخاذ رویکرد بردارهای متعامد توسط نویسندگان در روش پیشنهادی آن‌ها چیست و ایده کلی حمله به چه شکل است؟ (۵ نمره)
- تفاوت دو روش SimBA و SimBA-DCT در چیست؟ (۳ نمره)

با توجه به قطعی اینترنت و عدم امکان جست‌وجو تصویری برای درک بهتر DCT میتوانید از تصویر زیر استفاده کنید.



شکل ۱: Discrete Cosine Transform

- تفاوت دو روش SimBA و SimBA-DCT در چیست؟ (۳ نمره)
- بر اساس جدول ارزیابی روش‌ها در مقاله، حملات جعبه سیاه به دو دسته حملات امتیاز محور[۱۲] و حملات مبتنی بر برچسب[13] تقسیم شده‌اند. تفاوت اصلی بین این دو رویکرد را توضیح دهید. (۳ نمره)
- در جدول ارزیابی روش‌ها در مقاله، از ۳ معیار جهت ارزیابی روش‌ها استفاده شده است، هر یک را توضیح دهید. (۳ نمره)

---

[۱۲] label only
[13] score based

بخش سوم: پیاده سازی حملات جعبه سفید

در این بخش قصد داریم دو حمله جعبه سفید FGSM[14] و PGD[15] به صورت untargeted را پیاده سازی، برروی ResNet-20 که برمجموعه داده CIFAR-10 آموزش دیده است اعمال کنیم و تاثیر آن‌ها را مورد بررسی قرار خواهیم داد.

زیربخش اول: مدل و مجموعه داده

مجموعه داده CIFAR-10 شامل تصاویر ۳x۳۲x۳۲ در ۱۰ کلاس است. با توجه به محدودیت‌های اینترنت و منابع سخت‌افزاری، زیرمجموعه‌ای شامل ۲۵ تصویر که مدل آنها را به درستی دسته‌بندی کرده است، در اختیار شما قرار می‌گیرد. حملات متخاصم صرفا برروی این ۲۵ تصویر صورت می‌گیرد. وزن مدل آموزش دیده ResNet-20 هم در ایلرن قرار داده شده است.

برای بارگذاری مجموعه داده و مدل، از کد P2.ipynb استفاده کنید. لطفا در بخش Constants این نوت‌بوک، مسیر فایل‌های دانلود شده مجموعه داده و مدل را مشخص نمایید. برای استفاده از وزن‌های مدل، اطمینان حاصل کنید که پوشه pytorch_cifar_models در کنار نوت‌بوک P2.ipynb در یک دایرکتوری قرار گرفته باشد.

درصورت بارگذاری صحیح داده‌ها و مدل، باید در بخش Check Model and Dataset Load تمام تصاویر را مشاهده کنید و دقت مدل ٪۱۰۰ گزارش شود.

زیربخش دوم: FGSM (۱۲ نمره)

روش حمله FGSM را همراه با فرمول ریاضی آن توضیح دهید. سپس این حمله را پیاده‌سازی کرده و دقت مدل را پس از اعمال FGSM گزارش کنید. چهار نمونه از تصاویر خصمانه‌ای را که مدل در تشخیص آن‌ها دچار خطا می‌شود انتخاب کرده و برای هر نمونه موارد زیر را نمایش دهید: تصویر اصلی، تصویر خصمانه، نویز افزوده‌شده، برچسب واقعی و برچسب اشتباه پیش‌بینی‌شده.

مقدار پیشنهادی ε در فضای استانداردشده (با میانگین ۰ و واریانس ۱) برابر با ۰.۰۳ است.

[14] Fast Gradient Sign Method

[15] Projected Gradient Descent

زیربخش سوم: PGD (۱۲ نمره)

روش حمله PGD را همراه با فرمول ریاضی آن توضیح دهید. سپس این حمله را پیاده‌سازی کرده و دقت مدل را پس از اعمال PGD گزارش کنید. چهار نمونه از تصاویر خصمانه‌ای را که مدل در تشخیص آن‌ها دچار خطا می‌شود انتخاب کرده و برای هر نمونه موارد زیر را نمایش دهید: تصویر اصلی، تصویر خصمانه، نویز افزوده‌شده، برچسب واقعی و برچسب اشتباه پیش‌بینی‌شده.

در فضای استانداردشده (با میانگین ۰ و واریانس ۱)، مقدار پیشنهادی ε برای نورم بی‌نهایت برابر با ۰.۰۳، مقدار پیشنهادی α برابر با ۰.۰۰۵، و حداکثر تعداد گام‌های به‌روزرسانی برابر با ۲۰ است.

نکات تحویل

- مهلت ارسال این تمرین تا پایان روز "دوشنبه ۱۴ اردیبهشت ماه" خواهد بود.
- این زمان قابل تمدید نیست و درصورت نیاز می‌توانید از گریس استفاده کنید.
- در نظر داشته باشید که حداکثر مهلت آپلود تمرین در سامانه تا ۷ روز پس مهلت تحویل است و پس از آن سامانه بسته خواهد شد.
- در صورت وجود سوال و یا ابهام میتوانید از طریق ایمیل با موضوع TAI_HW1 با تدریسیاران آموزشی در ارتباط باشید. توجه کنید که حتما ایمیل سرتدریسیار نیز CC شود.
- حداکثر تعداد صفحات گزارش برای سوال اول: ۸ صفحه
- حداکثر تعداد صفحات گزارش برای سوال دوم: ۱۲ صفحه
- لطفا گزارش ، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه بارگذاری نمائید:

HW1 _[Lastname]_[StudentNumber].zip

- تدریسیاران تمرین:

۱. فرهاد نصری

f.nasri@ut.ac.ir

۲. مصطفی ابراهیمی

ebrahimi.mostafa@ut.ac.ir

- سرتدریسیار:

محمد گرجی

gorji.mohammad@ut.ac.ir

با آرزوی سلامتی و موفقیت روزافزون