به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس شبکه‌های عصبی و یادگیری عمیق

تمرین پنجم

پرسش ۱	نام دستیار طراح	آرین فیروزی
	رایانامه	arianfirooziM@gmail.com
پرسش ۲	نام دستیار طراح	محمد گرجی
	رایانامه	mohamadgorjicode@gmail.com
	مهلت ارسال پاسخ	۱۴۰۴.۰۳.۱۸

فهرست

شکل‌ها

قوانین

قبل از پاسخ دادن به پرسش‌ها، موارد زیر را با دقت مطالعه نمایید:

- از پاسخ‌های خود یک گزارش در قالبی که در صفحه‌ی درس در سامانه‌ی Elearn با نام ***REPORTS_TEMPLATE.docx*** قرار داده شده تهیه نمایید.
- پیشنهاد می‌شود تمرین‌ها را در قالب گروه‌های دو نفره انجام دهید. (بیش از دو نفر مجاز نیست و تحویل تک نفره نیز نمره‌ی اضافی ندارد) توجه نمایید الزامی در یکسان ماندن اعضای گروه تا انتهای ترم وجود ندارد. (یعنی، می‌توانید تمرین اول را با شخص A و تمرین دوم را با شخص B و ... انجام دهید)
- **کیفیت گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است**؛ بنابراین، لطفا تمامی نکات و فرض‌هایی را که در پیاده‌سازی‌ها و محاسبات خود در نظر می‌گیرید در گزارش ذکر کنید.
- در گزارش خود مطابق با آنچه در قالب نمونه قرار داده شده، برای شکل‌ها زیرنویس و برای جدول‌ها بالانویس در نظر بگیرید.
- الزامی به ارائه توضیح جزئیات کد در گزارش نیست، اما باید نتایج بدست آمده از آن را گزارش و تحلیل کنید.
- **تحلیل نتایج الزامی می‌باشد، حتی اگر در صورت پرسش اشاره‌ای به آن نشده باشد.**
- **دستیاران آموزشی ملزم به اجرا کردن کدهای شما نیستند**؛ بنابراین، هرگونه نتیجه و یا تحلیلی که در صورت پرسش از شما خواسته شده را به طور واضح و کامل در گزارش بیاورید. در صورت عدم رعایت این مورد، بدیهی است که از نمره تمرین کسر می‌شود.
- **کدها حتما باید در قالب نوت‌بوک با پسوند .ipynb تهیه شوند، در پایان کار، تمامی کد اجرا شود و خروجی هر سلول حتما در این فایل ارسالی شما ذخیره شده باشد.** بنابراین برای مثال اگر خروجی سلولی یک نمودار است که در گزارش آورده‌اید، این نمودار باید هم در گزارش هم در نوت‌بوک کدها وجود داشته باشد.
- **در صورت مشاهده‌ی تقلب امتیاز تمامی افراد شرکت‌کننده در آن، 100- لحاظ می‌شود.**
- تنها زبان برنامه نویسی مجاز **Python** است.
- **استفاده از کدهای آماده برای تمرین‌ها به هیچ وجه مجاز نیست. در صورتی که دو گروه از یک منبع مشترک استفاده کنند و کدهای مشابه تحویل دهند، تقلب محسوب می‌شود.**
- نحوه محاسبه تاخیر به این شکل است: پس از پایان رسیدن مهلت ارسال گزارش، حداکثر تا یک هفته امکان ارسال با تاخیر وجود دارد، پس از این یک هفته نمره آن تکلیف برای شما صفر خواهد شد.

- سه روز اول: بدون جریمه
- روز چهارم: ۵ درصد
- روز پنجم: ۱۰ درصد
- روز ششم: ۱۵ درصد
- روز هفتم: ۲۰ درصد

- حداکثر نمره‌ای که برای هر سوال می‌توان اخد کرد ۱۰۰ بوده و اگر مجموع بارم یک سوال بیشتر از ۱۰۰ باشد، در صورت اخد نمره بیشتر از ۱۰۰، اعمال نخواهد شد.
  - برای مثال: اگر نمره اخذ شده از سوال ۱ برابر ۱۰۵ و نمره سوال ۲ برابر ۹۵ باشد، نمره نهایی تمرین ۹۷.۵ خواهد بود و نه ۱۰۰.
- لطفا گزارش، کدها و سایر ضمایم را به در یک پوشه با نام زیر قرار داده و آن را فشرده سازید، سپس در سامانه‌ی Elearn بارگذاری نمایید:

HW[Number] _[Lastname]_[StudentNumber]_[Lastname]_[StudentNumber].zip

(مثال: HW1_Ahmadi_810199101_Bagheri_810199102.zip)

- برای گروه‌های دو نفره، بارگذاری تمرین از جانب یکی از اعضا کافی است ولی پیشنهاد می‌شود هر دو نفر بارگذاری نمایند.

پرسش ۱. طبقه بندی تصاویر با VIT

۱-۱. مقدمه (۱۰ نمره)

در کشاورزی، تشخیص به موقع بیماری ها و آفات در گیاهان، نقش مهمی در سلامت و باروری محصولات دارد. در این تمرین، می خواهیم دو نوع شبکه عصبی را به منظور طبقه بندی بیماری با استفاده از برگ گیاه استفاده کنیم:

ویژن ترنسفورمر (ViT): مدل هایی هستند که معماری ترنسفورمری و مکانیسم توجه را در دادگان تصویری استفاده میکنند.

شبکه عصبی کانوولوشنی (CNN): همانطور که قبلا با کارکرد این نوع شبکه های عصبی آشنا شدید، شبکه های عصبی کانوولوشنی ویژگی های تصویر را با استفاده از لایه های کانوولوشن استخراج میکنند.

هدف تمرین آشنایی با ساختار ویژن ترنسفورمر و تفاوت های آن با سایر شبکه های عصبی است. به این منظور از دادگان استفاده شده در مقاله استفاده خواهید کرد که برای راحتی کار از این لینک قابل دریافت است.

برای انجام تمرین نیاز است که این مقاله را مطالعه کنید.

در رابطه با مقاله و ویژن ترنسفورمر ها به سوالات زیر پاسخ دهید.

- با توجه به مقاله، اصلی ترین تفاوت و مزیت استفاده از مدل های ویژن ترنسفورمر در مقایسه با مدل های سنتی چیست؟ (۵ نمره)
- زمانی که دادگان موجود محدود است، کدام مدل بهتر عمل میکند؟ پاسخ خود را با توجه به ساختار و مکانیسم های به کار رفته در مدل ها توجیه کنید. (۵ نمره)

۱-۲. آماده سازی داده ها (۱۵ نمره)

داده ها را دریافت کرده و مراحل زیر را طی کنید:

- نمایش نمونه از هر کلاس: از هر کدام از ۱۰ کلاس موجود در دادگان یک تصویر را نشان دهید. (۲ نمره)
- بررسی توازن دادگان: تعداد تصویر های مربوط به هر کلاس را در یک جدول مقایسه کنید. آیا تعداد دادگان موجود متوازنند؟ در صورتی که پاسخنان منفی است، با توجه به داده ها

استدلال کنید که چه نوع تقویت داده ای میتوان استفاده کرد تا بدون آسیب جدی به کیفیت عکس ها داده ها را متعادل کرد و آن را اعمال کنید. (۷ نمره)

- در صورتی که پیش پردازش دیگری پیشنهاد میکنید که میتواند عملکرد را بهبود ببخشد، آن را روی داده ها اعمال کنید. با توجه به اندازه ی ورودی مدل ها و اندازه ی پیشنهاد شده در مقاله، عکس ها باید در ابعاد خاصی باشند و به این منظور بایستی ساختار داخلی CNN مورد استفاده (که Inception-V3 است) باید تغییر کند، اما با توجه به اینکه موضوع اصلی تمرین ویژن ترنسفورمر است، میتوانید از اندازه ی دیگری استفاده کنید. (۳ نمره)
- داده های خود را به دو دسته ی آموزش و اعتبارسنجی تقسیم کنید. (۳ نمره)

۱-۳. آموزش مدل CNN (۲۰ نمره)

ابتدا مدل Inception-V3 را به صورت خام و بدون آموزش اولیه بارگذاری کنید. تعداد خروجی را با توجه به دادگان تنظیم کنید و سایر پارامتر ها را با توجه به مقاله مقدار دهی کنید. نحوه ی کارکرد کلی این مدل را شرح دهید. (۷ نمره)

تابع خطای استفاده در مقاله را توضیح دهید. چه توابع دیگری برای این کار مناسب است؟ (۳ نمره)

مدل را به مقدار حداقل ۱۰ دوره آموزش داده و نمودار دقت و خطا را برای هر دو دسته ی داده ها رسم کنید. نمودار آشفتگی مدل را رسم کنید. (۱۰ نمره)

۱-۴. آموزش مدل ViT (۴۰ نمره)

مدل شرح داده شده در مقاله را پیاده کنید. لازم است خروجی لایه های مدل را نمایش دهید و اگر در لایه ای بر خلاف آنچه در مقاله گفته شده عمل کردید، دلیل آن را ذکر کنید. (۲۰ نمره)

لایه Patch Embedding در شبکه های مبدل تصویر به چه منظوری استفاده میشود؟ کاهش یا افزایش اندازه ی هر patch در این تمرین چه تاثیراتی بر روی خروجی دارد؟ (۵ نمره)

(امتیازی) در لایه Patch Embedding یک قابلیت خروجی پیکسلی تعبیه کنید و به استفاده از آن خروجی این لایه را به صورت تصویر نشان دهید. (۵ نمره)

مدل را به مقدار حداقل ۲۰ دوره آموزش داده و نمودار دقت و خطا را برای هر دو دسته ی داده ها رسم کنید. نمودار آشفتگی مدل را رسم کنید. (۱۰ نمره)

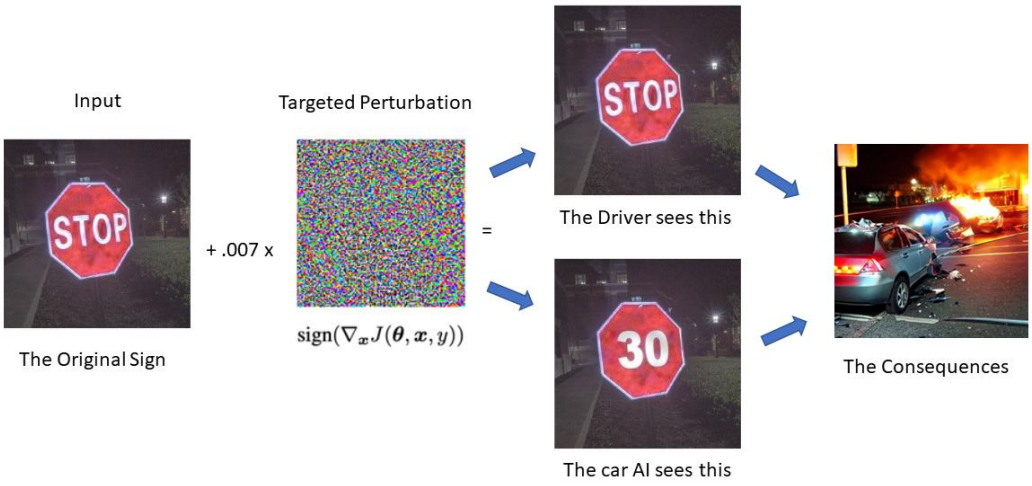
۱-۵. تحلیل و نتیجه گیری (۱۵ نمره)

نتیجه ی حاصل از دو روش را با همدیگر مقایسه کنید.

- مدل ها را از جوانب دقت، صحت و تعداد پارامتر با همدیگر مقایسه کنید. (۸ نمره)
- به توجه به اینکه کدام مدل بهتر عمل کرده، شرح دهید در چه شرایطی مدل ضعیفتر میتوانست بهتر عمل کند. (۷ نمره)

پرسش ۲ – Robust Zero-Shot Classification

حملات خصمانه در شبکه‌های عصبی یکی از چالش‌های جذاب و بحث‌برانگیز در یادگیری عمیق و هوش مصنوعی قابل اعتماد است. در این حملات، نمونه‌هایی طراحی می‌شوند که از نظر انسان عملاً غیرقابل تمایز با نمونه‌های اصلی هستند اما می‌توانند مدل‌های یادگیری ماشین را فریب دهند. این نمونه‌ها که به آن‌ها **نمونه‌های خصمانه**[۱] گفته می‌شود، با تغییرات کوچک و هدفمند در داده‌های ورودی ایجاد می‌شوند و باعث می‌شوند که مدل به خروجی نادرستی برسد. به عنوان مثال، افزودن نویزهای جزئی به یک تصویر می‌تواند باعث شود یک مدل طبقه‌بندی تصویر، تابلو ایست را به اشتباه به‌عنوان یک تابلو عبوری تشخیص دهد. (شکل ۱)



شکل ۱. اشتباه یک مدل هوش مصنوعی در تشخیص کلاس یک نمونه خصمانه

[۱] Adversarial Example

۲-۱. آشنایی با مدل CLIP، طبقه بندی تک ضرب و حملات خصمانه

ابتدا مقاله را مطالعه کرده و به سوالات زیر پاسخ دهید:

۱. دو روش تولید نمونه های تخاصمی FGSM و PGD را با شرح جزئیات بهینه سازی توضیح دهید؟ (۶ نمره)

۲. معماری مدل CLIP را همراه با شکل شرح داده و در مورد تابع زیان[۱] و نوع آموزش Contrastive توضیح دهید. (۶ نمره)

۳. تفاوت‌های کلیدی بین طبقه بندی عادی و طبقه بندی تک ضرب[۲] را توضیح دهید و همراه با شکل بیان کنید این عمل چگونه با مدل CLIP انجام می‌پذیرد. (۶ نمره)

۴. در شرایط مختلف دو نوع حمله خصمانه جعبه سفید و جعبه سیاه[۳] وجود دارد. هر کدام را توضیح دهید و سپس این دو رویکرد را از جنبه های مختلف دلخواه با یکدیگر مقایسه کنید. (۶ نمره)

۵. بیان بدارید چرا حملات انتقالی[۴] یک تهدید جدی برای مسائل دنیای واقعی به نسبت حملات جعبه سفید محسوب می‌شوند. (۶ نمره)

۶. روش تنظیم دقیق[۵] LoRA را همراه با شکل شرح دهید و سه علت استفاده از این روش نسبت به دیگر روش ها را بیان کنید و توضیح دهید؟ (۱۰ نمره)

۷. دو مقاله پژوهشی که تابع زیان CLIP را گسترش یا بهبود می‌دهند پیدا کنید و هر کدام را در یک الی دو پاراگراف خلاصه کنید. همچنین توضیح دهید که تاثیر این توابع زیان بر افزایش مقاومت مدل CLIP چگونه بوده است. (۱۰ نمره)

---

۱ Loss Function

۲ Zero-shot Classification

۳ White-box & Black-box

۴ Transfer Attacks

۵ Fine-tuning

۲-۲. پیاده سازی و مقایسه روش های آموزش خصمانه

در بخش دوم این پروژه، شما قرار است با بهره‌گیری از مجموعه داده CIFAR-10 و کتابخانه‌های PyTorch، torchvision، Transformers و PEFT، فرآیند آموزش و ارزیابی مدل‌های تک ضرب CLIP را در مقابل حملات خصمانه به صورت کامل پیاده‌سازی کنید.

۱. برای این سوال بایستی از دادگان CIFAR-10 استفاده کنید. این داده‌ها را با استفاده از ماژول torchvision دانلود کرده و به سه مجموعه آموزش، اعتبار سنجی و آزمایش تقسیم کنید. از یک تابع برای نمایش ۵ نمونه تصادفی از دادگان استفاده کنید سپس تصاویر را با تابع‌های تبدیل مناسب به سایز ۲۲۴×۲۲۴ تغییر دهید و با میانگین و انحراف معیار مختص CLIP نرمال‌سازی کنید. (۵ نمره)

۲. پس از آماده‌سازی داده‌ها، مدل CLIP را از روی مخزن HuggingFace بارگذاری کرده و در حالت ارزیابی قرار دهید. همچنین، به عنوان مدل مولد نمونه های خصمانه، یک مدل ResNet-20 پیش‌آموزش‌دیده روی CIFAR-10 را همانند تکه کد زیر از طریق PyTorch Hub دانلود کنید و آن را نیز در حالت eval نگه دارید. در ادامه، با تولید بردارهای متنی برای هر یک از ده کلاس CIFAR-10 (مثلاً a photo of a airplane، a photo of a automobile و...) فضای متنی CLIP را آماده کنید تا در مراحل بعدی از آن برای طبقه‌بندی تک ضرب بهره ببرید. (۵ نمره)

```
import torch
target_model = torch.hub.load("chenyaofo/pytorch-cifar-models",
"cifar10_resnet20", pretrained=True)
target_model.eval()
```

۳. مرحله بعدی ارزیابی مدل CLIP روی تصاویر تمیز است: با نوشتن تابعی که بردار ویژگی‌های تصویر را محاسبه و سپس نرمال‌سازی می‌کند، و با ضرب داخلی این بردارها در بردارهای متنی، خروجی طبقه‌بندی را به دست آورید و دقت مدل را گزارش کنید. سپس با استفاده از یک حمله PGD که روی مدل ResNet-20 انجام می‌شود (ε=8/255, α=2/255, steps=7)، نمونه‌های خصمانه بسازید (با استفاده از کتابخانه torchattacks) و همان تابع ارزیابی را اجرا کنید تا دقت CLIP در مواجهه با این حملات انتقالی را مشاهده نمایید. برای ملموس‌تر شدن موضوع، یک مثال تصویری از یک تصویر آزمون را بردارید و در کنار نسخه خصمانه‌اش و خود نویز حمله، در یک شکل سه‌تایی نمایش دهید. (۵ نمره)

۴. پس از آن، نوبت به تنظیم دقیق خصمانه معمولی با روش LoRA می‌رسد. برای این کار ابتدا با LoraConfig تنظیمات Low-Rank Adaptation را روی لایه‌های ماژول بینایی CLIP اعمال کنید (برای مثال r=8 و α=32). سپس مدل را به حالت train ببرید و در یک دوره آموزشی، برای هر بچ از تصاویر بردار ویژگی‌های CLIP را روی تصاویر خصمانه محاسبه و با بردارهای متنی ضرب داخلی کنید. با تابع CrossEntropyLoss روی برچسب‌های اصلی، گرادیان‌ها را محاسبه و پارامترهای LoRA را به‌روزرسانی نمایید. در پایان این مرحله، مجدداً دقت تمیز و خصمانه مدل را اندازه‌گیری کنید تا ببینید آموزش خصمانه استاندارد تا چه حد توانسته مقاومت مدل را افزایش دهد. (۱۵ نمره)

۵. . در گام بعد، الگوریتم TeCoA (Text-guided Contrastive Adversarial Training) را مطابق با معادله (۳) مقاله پیاده‌سازی کنید و مانند بخش قبل مدل را با این تابع هزینه آموزش داده و آزمایش های مربوطه را انجام دهید. (۱۵ نمره)

۶. در پایان، دقت‌های تمیز و خصمانه سه حالت اصلی مدل اولیه CLIP، CLIP تنظیم دقیق‌شده با LoRA و CrossEntropy، و CLIP تنظیم‌شده با LoRA و الگوریتم TeCoA را در یک جدول یا نمودار مقایسه‌ای نمایش دهید و به اختصار درباره مزایا و محدودیت‌های هر روش و میزان تقلیل یا افزایش دقت تمیز و مقاومت خصمانه بحث کنید. (۵ نمره)

۷. با استفاده از روش TeCoA مدل را بدون LoRA و با بکارگیری visual prompting tuning (مراجعه به مقاله اصلی) تنظیم دقیق کنید. در ادامه نیز این روش را با دو روش دیگر مقایسه و نتایج را تحلیل کنید. (۵ نمره امتیازی)

پ.ن.۱: مطابق با مقاله برای بخش های ۴ و ۵ شما بایستی روش های دو و پنج (ذکر شده در مقاله) را پیاده سازی و مقایسه کنید. البته که مطالعه دیگر روش ها نیز به فهم کلی مسئله کمک خواهد کرد:

(1) vanilla cross-entropy loss (CE)

(2) standard adversarial training loss (Adv.) with the cross-entropy loss

(3) contrastive adversarial training loss (CoAdv.)

(4) contrastive adversarial training over images (ImgCoAdv.)

(5) our text-guided contrastive adversarial training (TeCoA).

پ.ن.۲: برای آموزش از تعداد مساوی با تعداد نمونه های مجموعه داده آزمایشی استفاده کنید، نیاز به آموزش به روی کلیه مجموعه دادگان آموزشی نیست.