به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس شبکه‌های عصبی و یادگیری عمیق

پرسش ۱	نام و نام خانوادگی	علی قربانی برگانی
	شماره دانشجویی	۸۱۰۱۰۳۲۰۹
پرسش ۲	نام و نام خانوادگی	مبین تیرافکن
	شماره دانشجویی	۸۱۰۱۰۳۰۹۱
	مهلت ارسال پاسخ	

فهرست اصلی:

فهرست جداول:

فهرست تصاویر:

فهرست نمودار ها

پرسش۱ – طبقه بندی تصاویر با VIT

مقدمه

در دهه‌ی اخیر، یادگیری عمیق (Deep Learning) و به‌ویژه شبکه‌های عصبی کانولوشنی (CNNs) و ترنسفورمرها(Transformers) ، انقلابی شگرف در حوزه‌ی بینایی کامپیوتر (Computer Vision) و هوش مصنوعی پدید آورده‌اند. این مدل‌ها، با الهام از ساختار مغز انسان و توانایی یادگیری الگوهای پیچیده از حجم عظیم داده، به دستاوردهایی خارق‌العاده در اموری چون تشخیص اشیاء، تقسیم‌بندی تصاویر، تولید محتوای بصری و حتی تحلیل تصاویر پزشکی نائل آمده‌اند. توانایی این مدل‌ها در استخراج ویژگی‌های سلسله‌مراتبی از داده‌های خام، آن‌ها را به ابزاری بی‌بدیل در صنعت و پژوهش تبدیل کرده است.

اما این قدرت خیره‌کننده، بدون چالش نیست. عملکرد بالای این مدل‌ها اغلب در شرایط ایده‌آل و بر روی داده‌های تمیز و کنترل‌شده به دست می‌آید. دنیای واقعی، برخلاف دیتاست‌های آکادمیک، پر از نویز، عدم قطعیت و اختلالات پیش‌بینی‌نشده است. از سوی دیگر، ماهیت جعبه-سیاه (Black-Box) بسیاری از این معماری‌های پیچیده، درک رفتار و فرآیند تصمیم‌گیری آن‌ها را دشوار می‌سازد. این عدم شفافیت، یک سوال اساسی را مطرح می‌کند: "آیا می‌توانیم به تصمیمات این مدل‌ها، به خصوص در کاربردهای حیاتی مانند خودروهای خودران یا تشخیص‌های پزشکی، به طور کامل اعتماد کنیم؟"

در این گزارش جامع، ما سه جنبه‌ی کلیدی از عملکرد شبکه‌های عصبی مدرن را مورد کاوش قرار می‌دهیم:

۱. **تاثیر نویز بر فرآیند یادگیری :**در بخش نخست، به بررسی یکی از پایه‌ای‌ترین چالش‌های دنیای واقعی می‌پردازیم: کیفیت داده. ما یک مدل محبوب و قدرتمند به نام ResNet را یک بار بر روی داده‌های اصلی و تمیز دیتاست CIFAR-100 و بار دیگر بر روی همان داده‌ها پس از افزودن نویز گوسی آموزش خواهیم داد. با مقایسه‌ی دقیق منحنی‌های یادگیری و دقت نهایی، به این پرسش پاسخ می‌دهیم که چگونه وجود اختلال در داده‌های آموزشی، توانایی مدل برای یادگیری الگوهای معنادار و تعمیم‌پذیری به داده‌های جدید را تحت تاثیر قرار می‌دهد.

۲. **قدرت انتقال یادگیری (Transfer Learning)** :آموزش مدل‌های عظیم از ابتدا، نیازمند منابع محاسباتی گسترده و دیتاست‌های بسیار بزرگ است. در بخش دوم، به سراغ یکی از کارآمدترین راهکارها در یادگیری عمیق، یعنی "انتقال یادگیری"، می‌رویم. ما یک معماری پیشرفته مبتنی بر ترنسفورمر، Vision Transformer (ViT)، را که قبلاً بر روی دیتاست عظیم ImageNet

آموزش دیده است، برای یک وظیفه‌ی جدید و تخصصی (تشخیص ۱۰۲ نوع گل مختلف) Fine-tuneمی‌کنیم. عملکرد این مدل را با یک مدل ViT دیگر که از صفر بر روی همین دیتاست آموزش داده شده، مقایسه کرده و نشان می‌دهیم که چگونه دانش انباشته شده از یک وظیفه، می‌تواند به شکل چشمگیری فرآیند یادگیری در وظیفه‌ای دیگر را تسریع و بهبود بخشد.

۳. **آسیب‌پذیری در برابر حملات متخاصم و راهکارهای دفاعی :**شاید شکننده‌ترین و نگران‌کننده‌ترین ویژگی مدل‌های یادگیری عمیق، آسیب‌پذیری آن‌ها در برابر "نمونه‌های متخاصم (Adversarial Examples) "باشد؛ ورودی‌هایی که با ایجاد تغییرات جزئی و نامحسوس برای انسان، می‌توانند مدل را به طور کامل فریب داده و منجر به تصمیمات فاجعه‌بار شوند. در بخش سوم، ما با پیاده‌سازی دو حمله‌ی استاندارد FGSM وPGD ، به طور سیستماتیک به تمام مدل‌های آموزش‌دیده‌ی خود حمله کرده و میزان افت دقت آن‌ها را می‌سنجیم. سپس، با استفاده از تکنیک دفاعی قدرتمند "آموزش متخاصم(Adversarial Training) "، مدل‌ها را در برابر این حملات مقاوم‌سازی کرده و با مقایسه‌ی نتایج قبل و بعد از دفاع، میزان اثربخشی این راهکار را به صورت کمی و کیفی تحلیل می‌کنیم.

این پروژه با استفاده از زبان برنامه‌نویسی پایتون و کتابخانه‌ی قدرتمند PyTorch بر روی پلتفرم Google Colabبا شتاب‌دهنده‌ی گرافیکی (GPU) پیاده‌سازی شده است. در طول این گزارش، تلاش بر این است که ضمن ارائه نتایج دقیق و نمودارهای گویا، تحلیل‌های عمیق و مفهومی از پدیده‌های مشاهده‌شده ارائه گردد تا در نهایت، درکی جامع از رفتار، قابلیت‌ها و محدودیت‌های شبکه‌های عصبی عمیق حاصل شود.

۱-۱ بررسی اثر نویز گوسی بر عملکرد مدلResNet

یکی از فروض بنیادین در بسیاری از مسائل یادگیری ماشین، دسترسی به داده‌های آموزشی پاک و باکیفیت است که به خوبی نمایانگر توزیع داده‌های واقعی باشند. اما در عمل، این فرض به ندرت برقرار است. داده‌هایی که در دنیای واقعی جمع‌آوری می‌شوند، تقریباً همیشه تحت تأثیر انواع مختلفی از نویز و اختلال قرار دارند. این نویز می‌تواند ناشی از منابع متعددی باشد: خطاهای سنسور دوربین در شرایط نوری نامناسب، فشرده‌سازی تصویر، خطاهای انتقال داده، یا حتی شرایط محیطی پیش‌بینی‌نشده.

نویز، اطلاعات نامربوط و گمراه‌کننده‌ای را به داده‌ها اضافه می‌کند که می‌تواند فرآیند یادگیری مدل را به شدت مختل سازد. مدلی که بر روی داده‌های نویزی آموزش می‌بیند، ممکن است به جای

یادگیری ویژگی‌های اصلی و معنادار اشیاء، شروع به یادگیری الگوهای آماری خود نویز کند. این پدیده منجر به دو مشکل اساسی می‌شود:

- **کاهش دقت (Underfitting)** :مدل قادر به یادگیری کامل الگوهای زیربنایی داده نیست و در نتیجه، حتی بر روی داده‌های آموزشی نیز به عملکرد مطلوبی نمی‌رسد.
- **کاهش تعمیم‌پذیری (Poor Generalization)** :حتی اگر مدل بتواند خود را به داده‌های آموزشی نویزی "بیش‌برازش (Overfit) "کند، در مواجهه با داده‌های تست که تمیز هستند یا نویز متفاوتی دارند، عملکرد بسیار ضعیفی از خود نشان خواهد داد.

در این بخش، ما به صورت کنترل‌شده و آزمایشگاهی، نویز گوسی (Gaussian Noise) را به عنوان یک مدل استاندارد از نویز تصادفی، به داده‌های آموزشی خود اضافه می‌کنیم. نویز گوسی به دلیل سادگی ریاضی و شباهت آن به بسیاری از انواع نویزهای طبیعی در دنیای فیزیکی، یک انتخاب متداول در پژوهش‌های این حوزه است. هدف ما مشاهده و اندازه‌گیری کمی تأثیر این نوع اختلال بر روی یک معماری استاندارد و قدرتمند، یعنیResNet ، است.

طراحی آزمایش

برای ارزیابی دقیق اثر نویز، یک آزمایش مقایسه‌ای به شرح زیر طراحی گردید:

- معماری مدل :از مدل ResNet-18استفاده شد ResNet . به دلیل استفاده از "اتصالات میان‌بر " (Shortcut Connections)، قادر به آموزش شبکه‌های بسیار عمیق‌تری نسبت به معماری‌های سنتی است و به شکل مؤثری مشکل "محو شدگی گرادیان (Vanishing Gradient) "را برطرف می‌کند. این ویژگی آن را به یکی از محبوب‌ترین و پراستفاده‌ترین مدل‌ها در بینایی کامپیوتر تبدیل کرده است. مدل بدون استفاده از هیچ‌گونه وزن از پیش آموزش‌دیده (pretrained=False)مقداردهی اولیه شد تا یادگیری صرفاً بر اساس دیتاست مورد نظر صورت گیرد.
- دیتاست :دیتاست **CIFAR-100**برای این آزمایش انتخاب شد. این دیتاست شامل ۶۰,۰۰۰ تصویر رنگی در ابعاد ۳۲×۳۲ پیکسل است که به ۱۰۲ کلاس مختلف تقسیم شده‌اند (۵۰,۰۰۰ تصویر برای آموزش و ۱۰,۰۰۰ برای تست). وجود ۱۰۲ کلاس، این دیتاست را به یک چالش معنادار برای طبقه‌بندی تبدیل می‌کند و آن را از دیتاست‌های ساده‌تر مانند CIFAR-10 متمایز می‌سازد.

- تزریق نویز :یک تابع تبدیل (Transform) سفارشی طراحی شد تا به هر تصویر در مجموعه داده آموزشی، نویز گوسی با میانگین $\mu=0$ و انحراف معیار $\sigma=0.05$ اضافه کند. این عملیات فقط بر روی داده‌های آموزشی اعمال شد. مجموعه داده تست دست‌نخورده و تمیز باقی ماند تا بتوانیم توانایی تعمیم‌پذیری هر دو مدل را در شرایطی واقع‌گرایانه ارزیابی کنیم.
- سناریوهای آموزش :دو سناریوی مجزا پیاده‌سازی شد:

۱. مدل پایه (ResNet-Clean) :مدل ResNet-18 بر روی داده‌های آموزشی اصلی و بدون نویز CIFAR-100 آموزش داده شد.

۲. مدل نویزی :(ResNet-Noisy) همان مدل ResNet-18 بر روی داده‌های آموزشی که به آن‌ها نویز گوسی اضافه شده بود، آموزش داده شد.

- هایپرپارامترها :برای هر دو سناریو، هایپرپارامترهای یکسانی جهت اطمینان از صحت مقایسه به کار گرفته شد:
  - تعداد ایپاک (Epochs) :۲۰
  - بهینه‌ساز :Adam
  - تابع هزینه:Cross-Entropy Loss

نتایج و تحلیل عملکرد

پس از اتمام فرآیند آموزش برای هر دو سناریو، نتایج کمی و کیفی به دست آمده به دقت مورد بررسی قرار گرفت. در این بخش، ابتدا نتایج عددی ارائه شده و سپس با استفاده از نمودارهای یادگیری، تحلیل عمیقی از رفتار دو مدل ارائه می‌گردد.

۱-۳-۱ -نتایج کمی

لاگ‌های خروجی فرآیند آموزش، داده‌های ارزشمندی را در مورد عملکرد لحظه‌ای مدل‌ها در اختیار ما قرار می‌دهند. مهم‌ترین شاخص‌ها در پایان ۲۰ ایپاک در جدول زیر خلاصه شده‌اند:

شاخص عملکرد	ResNet-Clean	ResNet-Noisy
بیشترین دقت اعتبارسنجی	44.22%	33.69%
دقت اعتبارسنجی نهایی	43.30%	32.45%
خطای آموزشی نهایی	0.1676	0.5623
خطای اعتبارسنجی نهایی	3.9886	4.0254

جدول ۱ - مقایسه آموزش با داده های نویزی و بدون نویز برای **ResNet**

از جدول بالا چند نکته کلیدی قابل استخراج است:

۱. **شکاف عملکردی بزرگ :**یک اختلاف عملکرد فاحش در حدود ۱۰ الی ۱۱ درصد در بیشترین دقت اعتبارسنجی بین دو مدل وجود دارد. این موضوع به وضوح نشان می‌دهد که وجود نویز گوسی تا چه حد توانایی مدل برای یادگیری و تعمیم را کاهش داده است.

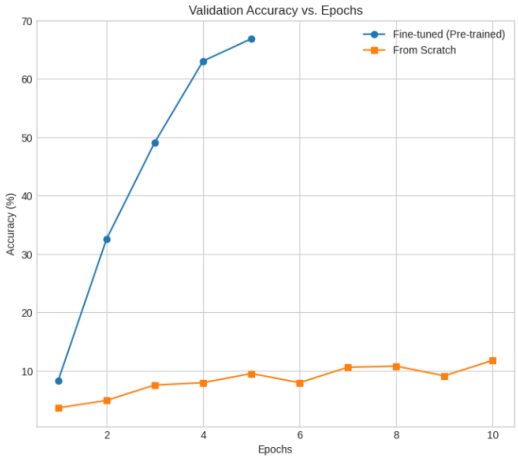
۲. **سرعت یادگیری :**مدل ResNet-Cleanبسیار سریع‌تر به دقت‌های بالا دست می‌یابد. این مدل تنها در ایپاک سوم به دقت ۳۶.۸۱٪ می‌رسد، در حالی که مدل ResNet-Noisyدر انتهای فرآیند آموزش (ایپاک ۱۹) به بهترین دقت خود یعنی ۳۳.۶۹٪ دست پیدا می‌کند.

۳. **میزان یادگیری داده‌های آموزشی :**خطای آموزشی نهایی در مدل ResNet-Clean (0.1676)به مراتب کمتر از مدل ResNet-Noisy (0.5623)است. این نشان می‌دهد که مدل اول توانسته است داده‌های آموزشی تمیز را بسیار بهتر و با جزئیات بیشتری یاد بگیرد.

۲-۳-۱ -تحلیل بصری منحنی‌های یادگیری

نمودار ارسالی شما که در زیر نیز آورده شده است، داستان کامل‌تری از فرآیند یادگیری دو مدل را روایت می‌کند.

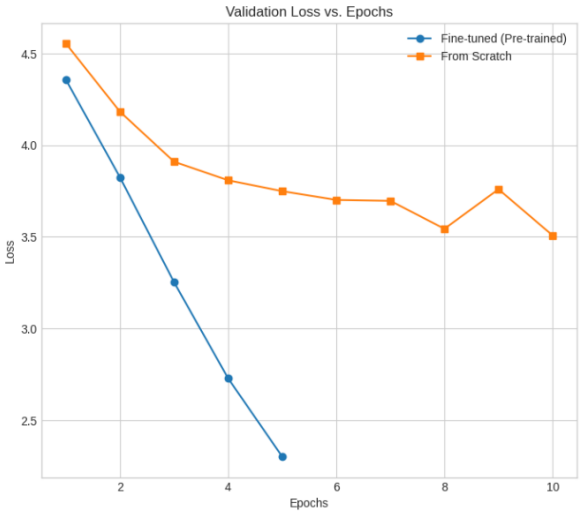
تحلیل نمودار دقت اعتبارسنجی(Validation Accuracy) :



نمودار ۱- نمودار **validation accuracy** مربوط به **ResNet**

- منحنی مدل ResNet-Clean (آبی):این منحنی یک رفتار یادگیری کلاسیک و سالم را نشان می‌دهد. دقت به سرعت افزایش یافته و در حوالی ایپاک هفتم به یک پلاتو (ناحیه صاف) در حدود ٪۴۴ می‌رسد. این رشد سریع نشان‌دهنده آن است که مدل به راحتی قادر به استخراج ویژگی‌های معنادار از تصاویر تمیز و یادگیری الگوهای مرتبط با ۱۰۰ کلاس مختلف است. پس از رسیدن به قله، دقت دچار نوسانات جزئی شده و کمی کاهش می‌یابد که می‌تواند از نشانه‌های شروع بیش‌برازش (Overfitting) باشد.
- منحنی مدل ResNet-Noisy (نارنجی) :در مقابل، منحنی این مدل یک فرآیند یادگیری دشوار و پرچالش را به تصویر می‌کشد. سرعت رشد دقت بسیار کندتر است و منحنی دارای نوسانات بیشتری است. این رفتار نشان می‌دهد که مدل در تلاش است تا سیگنال (ویژگی‌های اصلی تصویر) را از نویز (اختلالات گوسی) تفکیک کند، اما در این کار با دشواری مواجه است. بهترین عملکرد این مدل (حدود ٪۳۳) به طور قابل توجهی پایین‌تر از مدل پایه است که نشان‌دهنده سقف عملکردی پایین‌تر در حضور نویز است.

تحلیل نمودار خطای اعتبارسنجی(Validation Loss)



نمودار ۲ - نمودار **validation loss** مربوط به **Resnet**

- **منحنی مدل ResNet-Clean :**خطای اعتبارسنجی این مدل به سرعت کاهش یافته و در ایپاک هفتم به کمترین مقدار خود (حدود ۲.۲۶) می‌رسد .نکته بسیار مهم در این منحنی، رفتار آن پس از این نقطه است. مشاهده می‌کنیم که با ادامه فرآیند آموزش، خطای اعتبارسنجی به طور پیوسته و با شیب تندی افزایش می‌یابد. این یک علامت کلاسیک و قطعی از بیش‌برازش (Overfitting)است. در حالی که مدل در حال کاهش خطای خود بر روی داده‌های آموزشی است) همانطور که از ستون Train Loss در لاگ‌ها پیداست(، عملکرد آن بر روی داده‌های جدید (مجموعه اعتبارسنجی) در حال بدتر شدن است. مدل به جای یادگیری الگوهای عمومی، شروع به حفظ کردن جزئیات و نویزهای جزئی موجود در داده‌های آموزشی کرده است.
- **منحنی مدل ResNet-Noisy:** خطای اعتبارسنجی این مدل نیز کاهش می‌یابد اما هرگز به سطح پایین خطای مدل تمیز نمی‌رسد و کمترین مقدار آن (حدود ۳.۰۱ در ایپاک ۹) به طور قابل توجهی بالاتر است. اگرچه این منحنی نیز پس از مدتی روند صعودی به خود می‌گیرد، اما شیب افزایش آن به تندی مدل تمیز نیست.

آزمایش انجام شده در این بخش به طور قاطع نشان داد که کیفیت داده‌های آموزشی، یک فاکتور حیاتی و تعیین‌کننده در موفقیت یک مدل یادگیری عمیق است.

حضور نویز گوسی در تصاویر آموزشی، همانند یک مه غلیظ عمل می‌کند که دید مدل را برای تشخیص ویژگی‌های کلیدی تار می‌کند. این امر منجر به موارد زیر شد:

۱. **کاهش چشمگیر دقت نهایی :**مدل در بهترین حالت خود حدود ٪۱۱ دقت کمتری نسبت به زمانی داشت که روی داده‌های تمیز آموزش دیده بود.

۲. **کند شدن فرآیند یادگیری :**مدل برای رسیدن به بهترین عملکرد خود به زمان و تکرار بیشتری نیاز داشت.

۳. **دشواری در همگرایی :**مدل نویزی هرگز نتوانست به سطح پایین خطا که مدل تمیز به آن دست یافته بود، برسد که نشان‌دهنده دشواری بهینه‌ساز در یافتن یک مینیمم مناسب در فضای پارامتری است.

جالب است که پدیده بیش‌برازش در مدل آموزش‌دیده روی داده‌های تمیز بسیار واضح‌تر و شدیدتر بود. این به آن دلیل است که مدل ResNet-18 به قدری قدرتمند است که به سرعت الگوهای موجود در داده‌های آموزشی تمیز را یاد گرفته و سپس به سمت حفظ کردن آن‌ها حرکت می‌کند. در مقابل، در مدل نویزی، خود نویز به نوعی نقش یک منظم‌ساز (Regularizer) را ایفا می‌کند. این نویز تصادفی مانع از آن می‌شود که مدل بتواند به راحتی داده‌های آموزشی را حفظ کند. اما باید توجه داشت که این "اثر مثبت" در کاهش بیش‌برازش، در مقابل آسیب بسیار بزرگ‌تری که به توانایی یادگیری کلی مدل وارد می‌کند، کاملاً ناچیز و بی‌اهمیت است.

در نهایت، این بخش اهمیت مراحل پیش‌پردازش داده (Data Preprocessing) و تلاش برای اطمینان از کیفیت بالای داده‌های ورودی به مدل‌های یادگیری عمیق را برجسته می‌سازد.

۱-۲ انتقال یادگیری با ViT روی Flowers-102

در بخش اول، ما یک مدل ResNet-18 را از ابتدا (from scratch) بر روی دیتاست CIFAR-100 آموزش دادیم. اگرچه این یک تمرین آموزشی عالی است، اما در بسیاری از کاربردهای عملی در دنیای واقعی، این رویکرد نه بهینه و نه ممکن است. معماری‌های مدرن یادگیری عمیق، مانند خانواده ResNet یا مدل‌های جدیدتر مانندVision Transformer ، دارای ده‌ها و حتی صدها میلیون پارامتر قابل یادگیری هستند. آموزش موفقیت‌آمیز چنین شبکه‌های عظیمی نیازمند دو منبع گران‌بها است:

۱. **داده‌های عظیم :**دیتاست‌هایی در مقیاس میلیون‌ها تصویر مانند ImageNet برای جلوگیری از بیش‌برازش و یادگیری الگوهای عمومی ضروری هستند.

۲. **توان محاسباتی بسیار بالا :**آموزش این مدل‌ها نیازمند هفته‌ها یا حتی ماه‌ها پردازش بر روی خوشه‌هایی از واحدهای پردازش گرافیکی (GPU) یا تنسوری (TPU) است.

اینجاست که **انتقال یادگیری** به عنوان یک راهکار فوق‌العاده کارآمد وارد میدان می‌شود. ایده اصلی در انتقال یادگیری بسیار شهودی و مشابه فرآیند یادگیری انسان است: ما از دانش کسب‌شده در یک زمینه برای کمک به یادگیری سریع‌تر و بهتر در زمینه‌ای جدید اما مرتبط استفاده می‌کنیم. در بینایی

کامپیوتر، این ایده به این صورت ترجمه می‌شود: یک مدل بسیار بزرگ که قبلاً بر روی یک دیتاست عظیم و عمومی مانند ImageNet با ۱۰۰۰ کلاس مختلف از حیوانات و اشیاء روزمره آموزش دیده است، دانش بصری پایه‌ای و ارزشمندی را در لایه‌های خود ذخیره کرده است. این مدل یاد گرفته است که چگونه ویژگی‌های سطح پایین (مانند لبه‌ها، گوشه‌ها، رنگ‌ها و بافت‌ها) و ویژگی‌های سطح میانی (مانند اشکال، الگوها و قطعات اشیاء) را تشخیص دهد. این "دانش بصری" عمومی، تقریباً برای هر وظیفه بینایی کامپیوتر دیگری، از جمله تشخیص انواع گل، کاربرد دارد.

معماری مدل Vision Transformer (ViT) :ما از مدل ViTبه عنوان معماری اصلی خود استفاده می‌کنیم. این انتخاب تصادفی نیستViT .، که از موفقیت چشمگیر معماری ترنسفورمر در پردازش زبان طبیعی (NLP) الهام گرفته شده است، رویکردی متفاوت ازCNN های سنتی دارد. به جای استفاده از فیلترهای کانولوشنی، ViTیک تصویر را به مجموعه‌ای از تکه‌های (patches) مربعی تقسیم می‌کند، آن‌ها را به صورت خطی در بردارهایی جاسازی (embed) می‌کند و سپس این توالی از بردارها را به یک ترنسفورمر استاندارد می‌دهد. مکانیسم کلیدی در ترنسفورمر، "توجه به خود (Self-Attention) "است که به مدل اجازه می‌دهد تا ارتباط بین تمام تکه‌های تصویر را به صورت سراسری بسنجد و الگوهای پیچیده و دوربرد را تشخیص دهد. با این حال، ViTها به "داده-حریص (data-hungry) "بودن مشهور هستند و عملکرد آن‌ها در صورت آموزش از ابتدا بر روی دیتاست‌های کوچک، معمولاً ضعیف‌تر ازCNN هاست. این ویژگی، ViTرا به یک کاندیدای ایده‌آل برای نمایش قدرت انتقال یادگیری تبدیل می‌کند.

دیتاست هدف Flowers-102 :وظیفه ما طبقه‌بندی تصاویر در دیتاست Flowers-102است. این دیتاست چالش‌های خاص خود را دارد:

- تعداد کلاس بالا :شامل ۱۰۲ دسته مختلف از گل‌ها است.
- شباهت درون-کلاسی پایین و بین-کلاسی بالا :بسیاری از کلاس‌ها بسیار به یکدیگر شبیه هستند (مثلاً انواع مختلف گل رز) که این مسئله، طبقه‌بندی را به یک وظیفه "دقیق" یا "fine - grained"تبدیل می‌کند.
- **تعداد نمونه محدود :**تعداد تصاویر برای هر کلاس نسبتاً کم است (بین ۴۰ تا ۲۵۸ تصویر برای هر کلاس) که برای آموزش یک مدل پیچیده مانند ViT از ابتدا، بسیار ناکافی است.

دیتاست منبع ImageNet :مدل ViT ما با وزن‌های از پیش آموزش‌دیده بر روی دیتاست ImageNet (ILSVRC-2012)بارگذاری می‌شود. این دیتاست یک استاندارد طلایی در بینایی کامپیوتر است و شامل

بیش از ۱.۲ میلیون تصویر آموزشی در ۱۰۰۰ کلاس مختلف است. دانش استخراج شده از این دیتاست، یک پایه و اساس فوق‌العاده قوی برای تشخیص ویژگی‌های بصری فراهم می‌کند.

طراحی آزمایش

برای سنجش کمی و کیفی اثربخشی انتقال یادگیری، دو سناریوی آزمایشی مجزا مطابق با صورت تمرین طراحی و اجرا گردید:

سناریوی اول: تنظیم مجدد (Fine-tuning) مدل پیش‌آموزش‌دیده

۱. **بارگذاری مدل :**مدل vit_base_patch16_224با وزن‌های پیش‌آموزش‌دیده بر روی ImageNet-1Kبارگذاری شد. عبارت baseبه اندازه مدل، patch16به اندازه تکه‌های تصویر (۱۶×۱۶ پیکسل) و 224به ابعاد تصویر ورودی (۲۲۴×۲۲۴ پیکسل) اشاره دارد.

۲. **تطبیق لایه خروجی :**لایه طبقه‌بندی نهایی مدل (که heads.headنام دارد) برای خروجی ۱۰۰۰ کلاس ImageNet طراحی شده بود. این لایه حذف و با یک لایه خطی جدید جایگزین شد که دارای ۱۰۲ نورون خروجی، متناسب با تعداد کلاس‌های دیتاستFlowers-102 ، است.

۳. **فرآیند تنظیم مجدد (Fine-tuning):** تمام پارامترهای مدل، از جمله لایه‌های استخراج ویژگی و لایه طبقه‌بندی جدید، برای ۵ ایپاک بر روی دیتاست Flowers-102 آموزش داده شدند. استفاده از تعداد ایپاک کم در این سناریو منطقی است، زیرا مدل تنها نیاز به "تنظیم" و "تطبیق" دانش قبلی خود با داده‌های جدید دارد و نیازی به یادگیری از صفر نیست. یک نرخ یادگیری پایین معمولاً برای این مرحله انتخاب می‌شود تا دانش ارزشمند ذخیره شده در لایه‌های اولیه از بین نرود.

سناریوی دوم: آموزش مدل از ابتدا(Training from Scratch)

۱. **بارگذاری مدل :**همان معماری vit_base_patch16_224بارگذاری شد، اما این بار بدون وزن‌های پیش‌آموزش‌دیده. پارامترهای مدل به صورت تصادفی مقداردهی اولیه شدند. لایه خروجی نیز از ابتدا برای ۱۰۲ کلاس تنظیم شد.

۲. **فرآیند آموزش :**این مدل برای ۱۰ ایپاک بر روی دیتاست Flowers-102 آموزش داده شد. تعداد ایپاک‌های بیشتر (دو برابر سناریوی اول) به این دلیل انتخاب شد تا به مدل فرصت کافی برای یادگیری ویژگی‌ها از صفر داده شود، هرچند انتظار می‌رود که حتی با این تعداد ایپاک نیز مدل در یادگیری با چالش مواجه شود.

در هر دو سناریو، از هایپرپارامترهای مشابهی برای بهینه‌ساز (Adam) و تابع هزینه (Cross-Entropy Loss) استفاده شد تا مقایسه تا حد امکان منصفانه باشد.

نتایج و تحلیل عملکرد

نتایج به دست آمده از اجرای دو سناریوی "تنظیم مجدد" و "آموزش از ابتدا" به طرز چشمگیری متفاوت بودند و به وضوح قدرت پارادایم انتقال یادگیری را به تصویر می‌کشند.

جدول زیر، خلاصه عملکرد نهایی دو مدل را بر اساس لاگ‌های خروجی ارائه می‌دهد:

شاخص عملکرد	سناریو ۱ ViT Fine-tuned	سناریو ۲ ViT from Scratch
تعداد ایپاک آموزش	۵	۱۰
دقت اعتبارسنجی نهایی	66.86%	11.76%
خطای اعتبارسنجی نهایی	2.3016	3.5072
زمان کل آموزش	~۴ دقیقه(3m 60s)	~۸ دقیقه(8m 1s)

جدول ۲ - خلاصه عملکرد نهایی دو مدل **Fine tuned** و **Scratch**

تحلیل اعداد خام در این جدول، خود گویای همه چیز است:

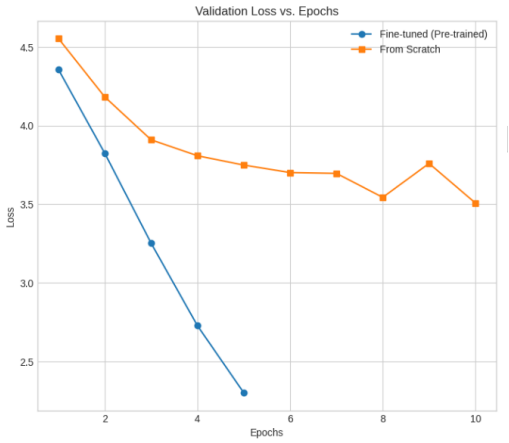
۱. **اختلاف دقت فاحش :**مدل تنظیم مجدد شده به دقتی دست یافت که بیش از ۵.۵ برابر دقت مدل آموزش‌دیده از ابتدا است (۶۶.۸۶٪ در مقابل ۱۱.۷۶٪). این یک اختلاف عملکرد عظیم است.

۲. **کارایی زمانی و محاسباتی :**مدل Fine-tuned نه تنها به دقتی بسیار بالاتر رسید، بلکه این کار را در نصف زمان و با نصف تعداد ایپاک‌ها انجام داد. این موضوع اهمیت انتقال یادگیری را در پروژه‌های واقعی که با محدودیت زمان و منابع مواجه هستند، دوچندان می‌کند.

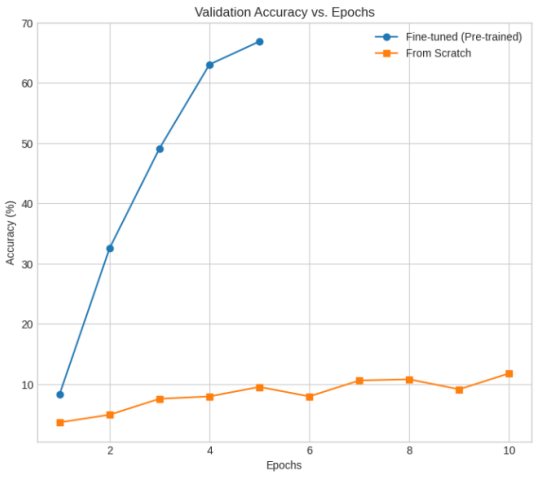
۳. **عملکرد مدل Scratch:** دقت ۱۱.۷۶٪ برای یک مسئله با ۱۰۲ کلاس، اگرچه کمی بهتر از حدس تصادفی (که حدود ۱٪ است) می‌باشد، اما در عمل یک عملکرد بسیار ضعیف و غیرقابل قبول محسوب می‌شود. این نشان می‌دهد که مدل در طول ۱۰ ایپاک آموزش، موفق به یادگیری الگوهای معنادار و قابل تعمیمی نشده است.

تحلیل بصری منحنی‌های یادگیری

تحلیل منحنی‌های مدلFine-tuned:



نمودار ۳ - نمودار **Validation Loss**مربوط به **Vit**



نمودار ۴ - نمودار **Validation Accuracy**مربوط به **ViT**

- این مدل یک شروع قدرتمند را تجربه می‌کند. حتی پس از ایپاک اول، دقت اعتبارسنجی آن به ۸.۳۳٪ می‌رسد که نشان می‌دهد دانش اولیه مدل از ImageNet بلافاصله به کار گرفته شده است.
- **رشد انفجاری دقت :**منحنی دقت با شیب بسیار تندی بالا می‌رود و به سرعت از مرزهای ۵۰٪ و ۶۰٪ عبور می‌کند. این نشان می‌دهد که فرآیند یادگیری بسیار کارآمد است. مدل نیازی به یادگیری ویژگی‌های پایه‌ای ندارد؛ بلکه تمام تمرکز خود را بر روی "تطبیق" دانش قبلی خود با ویژگی‌های ظریف و متمایزکننده گل‌ها معطوف می‌کند.

- **کاهش سریع خطا :**منحنی خطا نیز به طور متناظر، با سرعت بالایی کاهش می‌یابد که نشان‌دهنده همگرایی سریع و بهینه مدل به سمت یک راه حل خوب است.

تحلیل منحنی‌های مدل Scratch:

- این مدل یک فرآیند یادگیری بسیار کند، دشوار و طاقت‌فرسا را به نمایش می‌گذارد.
- **دقت نزدیک به صفر :**منحنی دقت (نارنجی) تقریباً بر روی محور افقی "می‌خزد". پس از ۵ ایپاک (کل زمان آموزش مدل اول)، دقت این مدل تنها به ۹.۵٪ رسیده است. حتی پس از ۱۰ ایپاک کامل، این منحنی به زحمت به بالای ۱۰٪ صعود می‌کند.
- **خطای بالا و کاهش ناچیز :**منحنی خطا (قرمز) در سطوح بسیار بالایی باقی می‌ماند و کاهش آن بسیار جزئی است. این نشان می‌دهد که مدل در یافتن الگوهای معنادار در داده‌ها کاملاً سردرگم است و بهینه‌ساز (Optimizer) نمی‌تواند گرادیان‌های مفیدی برای به‌روزرسانی وزنه‌ها پیدا کند.

۳-۳-۲ -تفسیر نتایج:

چرا چنین اختلاف عظیمی را شاهد هستیم؟ پاسخ در ماهیت دیتاست‌ها و معماری مدل نهفته است.

مدل ViT-Finetunedکار خود را از یک جایگاه بسیار ممتاز شروع کرد. لایه‌های اولیه این مدل، پس از پردازش بیش از یک میلیون تصویر متنوع ازImageNet ، به یک "کتابخانه بصری" غنی تبدیل شده بودند. این لایه‌ها می‌دانستند چگونه لبه‌ها، بافت‌ها، گرادیان‌های رنگی، و اشکال ابتدایی را تشخیص دهند. بنابراین، وقتی با تصاویر گل‌ها مواجه شدند، وظیفه‌شان یادگیری این مفاهیم پایه‌ای از صفر نبود؛ بلکه یادگیری ترکیب این مفاهیم برای تمایز قائل شدن بین کاسبرگ یک گل شقایق و گلبرگ یک گل لاله بود. این دقیقاً همان چیزی است که "تنظیم مجدد" یا Fine-tuning نامیده می‌شود.

در مقابل، مدل ViT-Scratchمانند یک نوزاد است. این مدل نه تنها باید یاد می‌گرفت که تفاوت‌های ظریف بین ۱۰۲ نوع گل چیست، بلکه باید مفاهیم بسیار ابتدایی‌تری مانند "گلبرگ چیست؟"، "ساقه چیست؟" و "بافت یک برگ چگونه است؟" را نیز از صفر می‌آموخت. دیتاست Flowers-102 با تعداد نمونه‌های محدود خود، به هیچ وجه داده‌های کافی برای آموزش این هرم دانش پیچیده از صفر را در اختیار یک مدل چند صد میلیون پارامتری مانند ViT قرار نمی‌دهد. در نتیجه، مدل در یک فضای پارامتری بسیار بزرگ سرگردان شده و قادر به یافتن یک راه حل معنادار نیست. این آزمایش به وضوح طبیعت "داده-حریص" بودن معماری‌های مبتنی بر ترنسفورمر را اثبات می‌کند.

این بخش از پروژه به صورت تجربی و با نتایجی قاطع، برتری مطلق استراتژی انتقال یادگیری را در سناریوهای عملی به اثبات رساند. نتایج نشان داد که انتقال یادگیری صرفاً یک "بهینه‌سازی" یا یک

"میان‌بر" نیست؛ بلکه برای استفاده مؤثر از معماری‌های مدرن و پیچیده مانند Vision Transformer بر روی دیتاست‌های تخصصی و با اندازه متوسط، یک ضرورت مطلق است.

مدل پیش‌آموزش‌دیده نه تنها به دقتی دست یافت که از نظر عملی قابل استفاده است (۶۶.۸۶٪)، بلکه این کار را با بهره‌وری محاسباتی بسیار بالاتری انجام داد. این یافته‌ها تاییدی بر این اصل مهم در یادگیری عمیق است که باید تا حد امکان از دانش انباشته شده در مدل‌های بزرگ که توسط جوامع تحقیقاتی و شرکت‌های پیشرو توسعه یافته‌اند، استفاده کرد و از تلاش برای "اختراع مجدد چرخ" در هر پروژه پرهیز نمود.

۱-۳ حملات متخاصم و دفاع

در بخش‌های پیشین، ما مدل‌هایی ساختیم که به دقت‌های قابل توجهی در وظایف طبقه‌بندی دست یافتند. این موفقیت‌ها ممکن است این تصور را ایجاد کند که این مدل‌ها درک عمیقی از جهان بصری پیدا کرده‌اند. اما این تصور، یک تصور شکننده است. پژوهش‌های گسترده در حوزه یادگیری ماشین متخاصم (Adversarial Machine Learning) نشان داده‌اند که مدل‌های یادگیری عمیق، علی‌رغم عملکرد فوق‌العاده‌شان، می‌توانند به طرز شگفت‌آوری فریب بخورند.

نمونه‌های متخاصم (Adversarial Examples) ورودی‌هایی هستند که به صورت عمدی و با ایجاد تغییراتی جزئی و هدفمند، برای فریب دادن یک مدل طراحی شده‌اند. این تغییرات اغلب برای چشم انسان کاملاً نامحسوس یا بی‌معنی هستند، اما می‌توانند مدل را وادار کنند که با اطمینان بالا، یک پیش‌بینی کاملاً اشتباه انجام دهد. برای مثال، یک تصویر از یک پاندا که مدل با اطمینان ۹۹٪ آن را تشخیص می‌دهد، می‌تواند با افزودن یک نویز هدفمند و بسیار خفیف، به تصویری تبدیل شود که مدل آن را با اطمینان ۹۹٪ به عنوان یک گیبون (نوعی میمون) طبقه‌بندی می‌کند، در حالی که برای انسان، تصویر همچنان یک پاندای بی‌نقص به نظر می‌رسد.

این پدیده صرفاً یک کنجکاوی آکادمیک نیست، بلکه دارای پیامدهای امنیتی بسیار جدی در دنیای واقعی است:

- در خودروهای خودران، یک برچسب کوچک و هوشمندانه طراحی‌شده بر روی یک تابلوی راهنمایی و رانندگی می‌تواند باعث شود که خودرو یک تابلوی "ایست" را به عنوان "محدودیت سرعت ۱۲۰ کیلومتر" تشخیص دهد.

- در سیستم‌های تشخیص چهره، یک عینک با طرح خاص می‌تواند هویت یک فرد را از دید سیستم پنهان کرده یا او را به عنوان فرد دیگری معرفی کند.
- در تشخیص‌های پزشکی، تغییرات جزئی در یک تصویر رادیولوژی می‌تواند منجر به تشخیص اشتباه یک تومور بدخیم به عنوان خوش‌خیم، یا برعکس شود.

بنابراین، درک، سنجش و افزایش مقاومت (Robustness) مدل‌ها در برابر این حملات، یک گام حیاتی برای ساختن سیستم‌های هوش مصنوعی قابل اعتماد و امن است.

حملات متخاصم به دو دسته کلی حملات جعبه-سفید (White-box) و جعبه-سیاه (Black-box) تقسیم می‌شوند. در حملات جعبه-سفید، مهاجم به تمام جزئیات مدل، از جمله معماری، پارامترها و وزنه‌ها، دسترسی کامل دارد. در این پروژه، ما از دو مورد از معروف‌ترین حملات جعبه-سفید استفاده می‌کنیم.

الف) روش علامت گرادیان سریع(Fast Gradient Sign Method - FGSM)

FGSMکه توسط گودفلو و همکارانش در سال ۲۰۱۴ معرفی شد، یک حمله ساده، سریع و در عین حال بسیار مؤثر است. ایده اصلی آن بسیار هوشمندانه است: برای اینکه بیشترین تغییر را در خروجی مدل ایجاد کنیم، باید ورودی را در جهتی تغییر دهیم که تابع هزینه (Loss Function) با بیشترین سرعت افزایش یابد. این جهت، دقیقاً همان جهت گرادیان تابع هزینه نسبت به تصویر ورودی است . FGSMیک گام بزرگ در این جهت برمی‌دارد. فرمول ریاضی آن به شرح زیر است:

$$x_{adv} = x + \epsilon \cdot sign\left(\nabla_x J(\theta, x, y)\right)$$

فرمول ۱-**FGSM**

در این فرمول:

- x_{adv}تصویر متخاصم تولید شده است.
- xتصویر اصلی و تمیز است.
- yبرچسب (کلاس) صحیح تصویر اصلی است.
- $J(\theta,x,y)$تابع هزینه مدل است که نشان می‌دهد پیش‌بینی مدل چقدر با برچسب واقعی فاصله دارد.
- ∇xJگرادیان تابع هزینه نسبت به پیکسل‌های تصویر ورودی xاست. این گرادیان به ما می‌گوید که با تغییر کدام پیکسل‌ها، خطا بیشتر می‌شود.

- sign(·)تابع علامت است که به هر عنصر از گرادیان، مقدار ۱، ۰ یا -۱ را نسبت می‌دهد. این کار باعث می‌شود که جهت کلی افزایش خطا را داشته باشیم.
- ϵ(اپسیلون) یک هایپرپارامتر کوچک است که اندازه یا قدرت حمله را کنترل می‌کند. مقادیر بزرگتر ϵباعث ایجاد تغییرات محسوس‌تر در تصویر، اما حملات قوی‌تر می‌شود. در این پروژه، از $\epsilon=0.1$استفاده شده است.

ب) نزول گرادیان (Projected Gradient Descent - PGD)

PGDرا می‌توان به عنوان نسخه قوی‌تر، تکرارشونده و هوشمندانه‌تر FGSM در نظر گرفت PGD .به جای برداشتن یک گام بزرگ، چندین گام کوچک در جهت افزایش خطا برمی‌دارد و به همین دلیل اغلب به عنوان یکی از قوی‌ترین حملات مرتبه اول شناخته می‌شود. دو ویژگی کلیدی PGD عبارتند از:

۱. **تکرار (Iteration) :**حمله در طی چندین مرحله (مثلاً ۷ تکرار در این پروژه) انجام می‌شود. در هر مرحله، یک گام کوچک به اندازه αدر جهت علامت گرادیان برداشته می‌شود.

۲. **پروژکشن (Projection) :**مهم‌ترین ویژگی PGD این است که پس از هر گام، تصویر تولید شده را بررسی می‌کند تا مطمئن شود که فاصله آن از تصویر اصلی xدر یک محدوده مشخص یک "توپ" با شعاع ϵباقی می‌ماند. اگر تغییرات از حد مجاز ϵفراتر رود، تصویر به مرز این محدوده "پروژکت" یا بازگردانده می‌شود. این کار تضمین می‌کند که حمله قدرتمند باقی بماند اما تغییرات کلی در تصویر همچنان محدود و نامحسوس باشد.

به طور خلاصه، PGDبا جستجوی دقیق‌تر در همسایگی تصویر اصلی، شانس بیشتری برای یافتن یک نمونه متخاصم مؤثر دارد و به همین دلیل، دفاع در برابر آن بسیار دشوارتر از FGSM است.

چگونه می‌توانیم مدل‌های خود را در برابر این حملات مقاوم کنیم؟ یکی از مؤثرترین و استانداردترین روش‌های دفاعی، آموزش متخاصم است.

در آموزش متخاصم، ما به جای آموزش مدل فقط بر روی داده‌های تمیز، فرآیند آموزش را با نمونه‌های متخاصم غنی می‌کنیم. به طور مشخص، در هر مرحله از آموزش:

۱. یک بچ (batch) از داده‌های آموزشی تمیز به مدل داده می‌شود.

۲. برای هر تصویر در بچ، یک نمونه متخاصم با استفاده از یک حمله قوی معمولاً PGD در لحظه (on-the-fly)تولید می‌شود.

۳. سپس مدل بر روی این بچ از داده‌های متخاصم آموزش داده می‌شود تا یاد بگیرد آن‌ها را به درستی طبقه‌بندی کند.

این فرآیند مدل را وادار می‌کند که صرفاً به ویژگی‌های سطحی و شکننده تکیه نکند، بلکه ویژگی‌های پایدارتر و معنادارتری را یاد بگیرد که حتی پس از اعمال اختلالات متخاصم نیز در تصویر باقی می‌مانند. در واقع، آموزش متخاصم باعث می‌شود که سطح تصمیم (decision boundary) مدل صاف‌تر و هموارتر شود و حساسیت آن به تغییرات کوچک در ورودی کاهش یابد.

یک پیامد مهم این روش، پدیده‌ای به نام **مبادله دقت Accuracy-Robustness Trade-off**است. اغلب، مدل‌هایی که به صورت متخاصم آموزش دیده‌اند، ممکن است دقت کمی پایین‌تری بر روی داده‌های تمیز نسبت به مدل‌های استاندارد داشته باشند، اما در عوض، دقت آن‌ها بر روی داده‌های متخاصم به طور چشمگیری بالاتر است.

طراحی آزمایش و پروتکل ارزیابی

برای یک ارزیابی جامع، ما تمام مدل‌های توسعه‌یافته در بخش‌های قبلی را تحت این پروتکل امنیتی قرار دادیم:

- مدل‌های تحت آزمایش:
  ۱. ResNet-Cleanآموزش‌دیده روی داده تمیز.CIFAR-100
  ۲. ResNet-Noisyآموزش‌دیده روی داده نویزی.CIFAR-100
  ۳. ViT-Finetunedتنظیم مجدد شده روی.Flowers-102
  ۴. ViT-Scratchآموزش‌دیده از ابتدا روی.Flowers-102
- پروتکل ارزیابی :دقت هر یک از این چهار مدل، و همچنین نسخه‌های دفاع‌شده‌ی آن‌ها، در سه حالت زیر سنجیده شد:
  ۱. حالت پایه :عملکرد بر روی داده‌های تست اصلی و تمیز.
  ۲. تحت حمله :عملکرد بر روی داده‌های تست پس از اعمال حملات FGSM و PGD با پارامترهای ذکر شده.
  ۳. پس از دفاع :عملکرد مدل‌های جدیدی که با استفاده از آموزش متخاصم برای ۱۰ ایپاک مقاوم‌سازی شده‌اند.

نتایج و تحلیل جامع امنیت مدل‌ها

پس از اجرای پروتکل ارزیابی امنیتی بر روی هر چهار مدل اصلی و همچنین نسخه‌های مقاوم‌سازی شده‌ی آن‌ها، نتایج کامل در جدول زیر گردآوری شده است. این جدول، سنگ بنای تحلیل ما برای درک آسیب‌پذیری و مقاومت مدل‌ها خواهد بود.

جدول نهایی مقایسه عملکرد مدل‌ها در سه حالت: پایه، تحت حمله و پس از دفاع

مدل	وضعیت	دقت روی داده تمیز	دقت تحت حمله FGSM	دقت تحت حمله PGD
ResNet-Clean	Original	43.30%	1.92%	0.23%
	Defended	**33.36%**	**5.81%**	**3.83%**
ResNet-Noisy	Original	32.45%	7.18%	4.85%
	Defended	**23.28%**	**3.80%**	**2.48%**
ViT-Finetuned	Original	62.81%	2.46%	0.016%
	Defended	**78.55%**	**23.06%**	**11.14%**
ViT-Scratch	Original	9.30%	4.28%	1.51%
	Defended	**14.91%**	**8.44%**	**3.55%**

جدول ۳ - مقایسه تمامی مدل ها از لحاظ دقت در برابر انواع حملات

این نتایج غنی، چندین داستان مهم و آموزنده را روایت می‌کنند که در ادامه به تفکیک به آن‌ها می‌پردازیم.

آسیب‌پذیری شدید مدل‌های استاندارد

اولین و واضح‌ترین نتیجه‌ای که از ردیف‌های "Original" جدول می‌توان گرفت، شکنندگی فوق‌العاده زیاد مدل‌هایی است که به صورت استاندارد آموزش دیده‌اند.

- فروپاشی کامل مدل‌های با عملکرد بالا :مدل ResNet-Cleanکه روی داده‌های تمیز به دقت قابل قبول ۴۳.۳۰٪ رسیده بود، تحت حملات FGSM و PGD تقریباً به طور کامل از کار می‌افتد و دقت آن به ترتیب به 1.92%و 0.23%سقوط می‌کند. این عملکرد حتی از حدس تصادفی نیز بدتر است و نشان می‌دهد مدل با اطمینان بالا در حال انجام پیش‌بینی‌های اشتباه است. داستان برای مدل ViT-Finetunedحتی فاجعه‌بارتر است. این مدل که با دقت ۶۲.۸۱٪ بهترین عملکرد را روی داده تمیز داشت، در برابر حمله قدرتمند PGD به دقت ناچیز 0.016%می‌رسد. این

یعنی از هر ۱۰,۰۰۰ تصویر متخاصم، شاید تنها یکی را درست تشخیص دهد! این نتایج به وضوح نشان می‌دهند که دقت بالا بر روی داده تمیز، هیچ تضمینی برای مقاومت مدل نیست.

"مقاومت تصادفی" در مدل‌های ضعیف‌تر

یک مشاهده جالب، مقایسه بین ResNet-Clean و ResNet-Noisy در حالت "Original" است.

- مدل ResNet-Noisy که به دلیل آموزش بر روی داده‌های نویزی، دقت پایین‌تری روی داده تمیز دارد (۳۲.۴۵٪)، در کمال تعجب مقاومت بیشتری در برابر حملات متخاصم از خود نشان می‌دهد) دقت ۷.۱۸٪ در برابر FGSM و ۴.۸۵٪ در برابر .(PGD این مقاومت بیشتر، از مدل ResNet-Clean که ظاهراً قوی‌تر بود، به دست آمده است .چرا؟ دلیل این پدیده آن است که آموزش با نویز گوسی، مدل را وادار کرده است که تا حدی اختلالات فرکانس بالا و بی‌معنی را نادیده بگیرد. از آنجایی که حملات متخاصم نیز اغلب نویزهایی با فرکانس بالا به تصویر اضافه می‌کنند، این "بی‌توجهی" آموخته‌شده، به صورت تصادفی به عنوان یک سپر دفاعی ضعیف عمل کرده است. این پدیده نشان می‌دهد که هر نوع منظم‌سازی (Regularization) که مدل را از تمرکز بیش از حد بر روی جزئیات دقیق باز دارد، می‌تواند به مقاومت کمک کند.

قدرت و هزینه دفاع: تحلیل آموزش متخاصم

این بخش، هسته اصلی تحلیل ما و نمایشگر واقعی تأثیر استراتژی دفاعی است. با مقایسه ردیف‌های "Original"و "Defended" به نتایج شگفت‌انگیزی می‌رسیم.

- مدل ViT-Finetuned ستاره نمایش دفاعی :این مدل برجسته‌ترین نمونه موفقیت است.
  - افزایش همزمان دقت و مقاومت :برخلاف انتظار "مبادله دقت-مقاومت"، در این مدل یک پدیده نادر و بسیار مطلوب رخ داده است. آموزش متخاصم نه تنها مقاومت مدل را به شدت افزایش داده، بلکه دقت آن بر روی داده‌های تمیز را نیز از ۶۲.۸۱٪ به 78.55% بهبود بخشیده است! این نشان می‌دهد که آموزش متخاصم به عنوان یک منظم‌ساز بسیار قوی عمل کرده و مدل را از یک "مینیمم محلی" نامناسب به یک نقطه بهینه بهتر در فضای پارامتری هدایت کرده که هم عمومی‌تر و هم مقاوم‌تر است.
  - جهش کوانتومی در مقاومت :دقت مدل در برابر حمله FGSM از ۲.۴۶٪ به 23.06% (نزدیک به ۱۰ برابر) و در برابر حمله قدرتمند PGD از صفر مطلق (۰.۰۱۶٪) به 11.14% (بیش از ۶۸۰ برابر) جهش کرده است. این یک بهبود خیره‌کننده است و نشان

می‌دهد که مدل دفاع‌شده اکنون قادر است در برابر حملاتی که قبلاً آن را فلج می‌کردند، ایستادگی کند.

- **مدل ResNet-Cleanنمایش کلاسیک مبادله دقت-مقاومت:**
  - هزینه مقاومت :همانطور که در تئوری انتظار می‌رود، دقت مدل دفاع‌شده بر روی داده تمیز از ۴۳.۳۰٪ به 33.36%کاهش یافته است. این همان هزینه یا "مالیاتی" است که برای کسب مقاومت پرداخت کرده‌ایم.
  - دستاوردهای مقاومت :در عوض، دقت مدل در برابر FGSM از ۱.۹۲٪ به 5.81%(۳ برابر) و در برابر PGD از ۰.۲۳٪ به 3.83%(نزدیک به ۱۷ برابر) افزایش یافته است. اگرچه این مقادیر هنوز پایین هستند، اما بهبود نسبی آن‌ها نشان می‌دهد که دفاع مؤثر بوده است.
- مدل‌های دیگر :مدل ViT-Scratchنیز پس از دفاع، هم در دقت تمیز و هم در مقاومت بهبود یافته است که مجدداً نقش آموزش متخاصم به عنوان یک منظم‌ساز را برای مدل‌های ضعیف‌تر تأیید می‌کند. جالب اینجاست که مدل ResNet-Noisyپس از دفاع، عملکرد ضعیف‌تری در همه زمینه‌ها از خود نشان داده که ممکن است به دلیل تداخل اثر منفی دو نوع اختلال (نویز گوسی و نویز متخاصم) در فرآیند آموزش باشد.

این بخش از پروژه، سفری عمیق به دنیای شکننده امنیت در یادگیری عمیق بود. نتایج به دست آمده چندین اصل کلیدی را به اثبات رساند:

۱. مدل‌های استاندارد، صرف نظر از معماری CNN یا Transformer و دقت اولیه‌شان، به شدت در برابر حملات متخاصم آسیب‌پذیر هستند و نمی‌توان به آن‌ها در کاربردهای حساس اعتماد کرد.

۲. حملات مبتنی بر گرادیان مانند FGSM و به خصوصPGD ، ابزارهای بسیار مؤثری برای کشف و سنجش این آسیب‌پذیری‌ها هستند.

۳. آموزش متخاصم یک استراتژی دفاعی قدرتمند و کارآمد است که قادر است مقاومت مدل‌ها را به طور چشمگیری افزایش دهد.

۴. مفهوم "مبادله دقت-مقاومت "یک واقعیت است، اما همیشه صادق نیست. همانطور که در مدل ViTمشاهده شد، گاهی اوقات دفاع می‌تواند منجر به یافتن مدل‌های بهتر در هر دو زمینه شود.

در نهایت، این تحلیل نشان می‌دهد که ارزیابی امنیتی و مقاوم‌سازی، بخش‌های جدایی‌ناپذیر و حیاتی از چرخه حیات توسعه مدل‌های هوش مصنوعی هستند و نباید به عنوان یک گزینه اختیاری به آن‌ها نگریسته شود.

نتیجه‌گیری نهایی کل پروژه

در این پروژه جامع، ما سه جنبه متفاوت اما حیاتی از عملکرد شبکه‌های عصبی عمیق را کاوش کردیم. در بخش اول، دیدیم که کیفیت داده تا چه حد بر فرآیند یادگیری تأثیرگذار است و چگونه نویز می‌تواند عملکرد مدل را به شدت تضعیف کند. در بخش دوم، قدرت و کارایی پارادایم انتقال یادگیری را به نمایش گذاشتیم و دیدیم که چگونه استفاده از دانش انباشته، به خصوص برای معماری‌های پیچیده‌ای مانندViT ، نه تنها مفید بلکه ضروری است. در نهایت، در بخش سوم، به دنیای تاریک امنیت مدل‌ها قدم گذاشتیم، شکنندگی آن‌ها را در برابر حملات متخاصم به اثبات رساندیم و نشان دادیم که چگونه با استفاده از تکنیک‌های دفاعی مانند آموزش متخاصم می‌توان سپری در برابر این آسیب‌پذیری‌ها ساخت.

این سه بخش در کنار هم، یک تصویر کامل‌تر از چالش‌ها و راهکارهای موجود در دنیای واقعی یادگیری عمیق ارائه می‌دهند. ساختن یک مدل هوش مصنوعی موفق، فراتر از دستیابی به یک عدد دقت بالا در یک جدول است؛ این فرآیند نیازمند درک عمیق از داده‌ها، بهره‌گیری هوشمندانه از منابع موجود، و توجه جدی به امنیت و قابلیت اطمینان مدل در شرایط پیش‌بینی‌نشده و متخاصم است.

۱-۴سوالات تئوری

پرسش ۱ : با وجود اینکه مدل ResNet دارای میلیون‌ها پارامتر است، چرا در بسیاری از موارد با داده‌های نویزی همچنان عملکرد مناسبی دارد؟

در نگاه اول، یک مدل با ظرفیت بالا (میلیون‌ها پارامتر) باید به شدت مستعد بیش‌برازش بر روی نویز باشد و کاملاً از کار بیفتد. اما همانطور که در بخش اول دیدیم، اگرچه عملکرد ResNetبا داده‌های نویزی به شدت افت کرد، اما به صفر نرسید و توانست تا حدی الگوها را یاد بگیرد (به دقت حدود ٪۳۳ رسید). این "مقاومت نسبی" ناشی از ویژگی‌های معماری هوشمندانه ResNetاست:

۱ .هسته اصلی: اتصالات پسماندی(Residual Connections)

ویژگی انقلابی ResNet، معرفی اتصالات میان‌بر یا پسماندی است. در یک شبکه عصبی ساده، هر لایه تلاش می‌کند تا نگاشت H(x)را مستقیماً یاد بگیرد. اما در ResNet، هر بلوک تلاش می‌کند تا تابع پسماند

$F(x)=H(x)-x$را یاد بگیرد، به طوری که خروجی نهایی $H(x)=F(x)+x$باشد. این تغییر به ظاهر ساده، دو اثر عمیق بر مقاومت در برابر نویز دارد:

- هموارسازی فضای هزینه (Smoother Loss Landscape) :اتصال مستقیم xبه خروجی، یک مسیر بدون مانع برای جریان گرادیان‌ها از لایه‌های انتهایی به ابتدایی فراهم می‌کند. این کار از مشکل "محو شدگی گرادیان" جلوگیری کرده و به بهینه‌ساز اجازه می‌دهد تا در یک فضای هزینه بسیار هموارتر حرکت کند. در حضور نویز، که باعث ایجاد نوسانات و مینیمم‌های محلی تیز در فضای هزینه می‌شود، این هموارى به مدل کمک می‌کند تا در این "تله‌ها" گیر نیفتد و به سمت یک راه حل عمومی‌تر حرکت کند.
- تمایل به یادگیری توابع ساده‌تر :اگر یک بلوک برای یادگیری یک ویژگی خاص ضروری نباشد، بهینه‌ساز به راحتی می‌تواند وزن‌های آن بلوک را به سمت صفر سوق دهد، به طوری که $F(x)$و در نتیجه $H(x)$. این یعنی بلوک به یک "اتصال همانی (Identity Mapping) "تبدیل می‌شود. در مقابل نویز، این ویژگی به مدل اجازه می‌دهد تا از یادگیری الگوهای پیچیده و نامرتبط نویز خودداری کند و یک تابع ساده‌تر و پایدارتر را یاد بگیرد.

۲ .اثر منظم‌سازی ضمنی(Implicit Regularization)

معماری ResNetبه صورت ضمنی دارای خواص منظم‌سازی است:

- رفتار شبه-گروهی (Ensemble-like Behavior) :به دلیل وجود مسیرهای متعدد ایجاد شده توسط اتصالات میان‌بر، یک ResNetعمیق را می‌توان به عنوان یک گروه (Ensemble) ضمنی از تعداد زیادی شبکه کم‌عمق‌تر در نظر گرفت. اطلاعات می‌توانند از مسیرهای کوتاه یا بلند به خروجی برسند. این تنوع در مسیرها، مانند ترکیب کردن پیش‌بینی چندین مدل مختلف عمل کرده و باعث می‌شود که مدل به ویژگی‌های خاص و شکننده کمتر تکیه کند و در نتیجه، در برابر نویز مقاوم‌تر باشد.
- نرمال‌سازی دسته‌ای (Batch Normalization) :هر بلوک ResNetبه طور استاندارد از لایه‌های Batch Normاستفاده می‌کند. این لایه‌ها با نرمال‌سازی فعال‌سازی‌ها در هر دسته کوچک، به پایداری فرآیند آموزش کمک شایانی می‌کنند و اثر منظم‌سازی ملایمی دارند. این پایداری به مدل کمک می‌کند تا با تغییرات آماری که نویز به داده‌ها تحمیل می‌کند، بهتر کنار بیاید.

بنابراین، ResNetنه به این دلیل که به نویز "ایمن" است، بلکه به این دلیل که معماری آن به طور ذاتی پایدارتر، هموارتر و دارای خاصیت گروهی ضمنی است، می‌تواند در حضور نویز عملکردی "مناسب" (و نه عالی) از خود به نمایش بگذارد.

پرسش ۲: مدل‌های Vision Transformer (ViT) در صورت استفاده از وزن‌های پیش‌آموزش‌دیده بهتر از حالت بدون pre-train عمل می‌کنند. این اختلاف را با استفاده از مفاهیم مینیمم تیز (sharp) و تخت (flat) تحلیل کنید.

نتایج بخش دوم پروژه ما به وضوح نشان داد که اختلاف عملکرد بین ViT-Finetunedو ViT-Scratch بسیار زیاد است. این پدیده را می‌توان با تحلیل شکل فضای هزینه (Loss Landscape) و نوع مینیمم‌هایی که هر مدل پیدا می‌کند، به زیبایی توضیح داد.

تعریف مینیمم‌های تخت و تیز:

- مینیمم تخت (Flat Minimum) :ناحیه‌ای وسیع و با انحنای کم در فضای هزینه است. مدل‌هایی که به این نوع مینیمم‌ها همگرا می‌شوند، تمایل به تعمیم‌پذیری بالا دارند. چرا؟ زیرا تغییرات کوچک در وزن‌های مدل (که ممکن است در اثر مواجهه با داده‌های تست رخ دهد) تأثیر کمی بر روی مقدار خطا دارد.
- مینیمم تیز (Sharp Minimum) :دره‌ای باریک و با انحنای زیاد در فضای هزینه است. مدل‌هایی که در این نقاط همگرا می‌شوند، به شدت بیش‌برازش (Overfit) شده‌اند و تعمیم‌پذیری ضعیفی دارند. در این حالت، مدل یک راه حل بسیار خاص برای داده‌های آموزشی پیدا کرده و کوچکترین تغییری در وزن‌ها، آن را از این نقطه بهینه خارج کرده و باعث افزایش شدید خطا می‌شود.

تحلیل دو سناریو:

۱. آموزش از ابتدا (ViT-Scratch) و مینیمم‌های تیز :وقتی یک مدل با ظرفیت بسیار بالا مانند ViT را بر روی یک دیتاست نسبتاً کوچک) مانند (Flowers-102 از ابتدا آموزش می‌دهیم، مدل به دلیل ظرفیت بالای خود، به راحتی می‌تواند داده‌های آموزشی را حفظ کند .این حفظ کردن، معادل پیدا کردن یک مینیمم تیز در فضای هزینه است. مدل یک پیکربندی بسیار خاص از میلیون‌ها پارامتر خود را پیدا می‌کند که خطای آموزشی را به حداقل می‌رساند، اما این پیکربندی بسیار شکننده است و نمایانگر توزیع واقعی داده‌ها نیست. به همین دلیل، با دیدن داده‌های جدید (تست)، عملکرد آن به شدت افت می‌کند.

۲. انتقال یادگیری (ViT-Finetuned) و مینیمم‌های تخت :مدل پیش‌آموزش‌دیده، سفر خود را از یک نقطه شروع فوق‌العاده در فضای پارامتری آغاز می‌کند. این نقطه وزن‌های حاصل از آموزش روی ImageNet قبلاً در یک ناحیه بسیار خوب از فضای هزینه قرار گرفته است؛ ناحیه‌ای که نمایانگر ویژگی‌های بصری عمومی و پایدار است. این "ناحیه خوب" ذاتاً یک مینیمم تخت است. ویژگی‌های یادگرفته شده از ImageNet آنقدر عمومی هستند که تغییرات کوچک در وزن‌ها، ماهیت آن‌ها را از بین نمی‌برد .فرآیندFine-tuning ، به جای جستجوی کل فضای پارامتری، تنها یک جستجوی محلی در این "دره تخت" انجام می‌دهد تا نقطه‌ای را پیدا کند که برای وظیفه جدید (تشخیص گل‌ها) کمی بهینه‌تر باشد. از آنجایی که کل این ناحیه تخت است، نقطه نهایی نیز در یک مینیمم تخت قرار خواهد گرفت که این امر منجر به تعمیم‌پذیری عالی و عملکرد فوق‌العاده مدل می‌شود.

به طور خلاصه، پیش‌آموزش، مدل را به سمت یک ناحیه امن و تخت هدایت می‌کند، در حالی که آموزش از ابتدا، مدل را در خطر افتادن در دره‌های تیز و بی‌ثبات قرار می‌دهد.

پرسش ۳ : Data Augmentationرا به عنوان شکلی از Adversarial Training تحلیل کنید.

در نگاه اول، Data Augmentation(مانند چرخش، برش، تغییر رنگ) و Adversarial Trainingدو تکنیک مجزا به نظر می‌رسند. اما در هسته، هر دو یک هدف مشترک را دنبال می‌کنند :افزایش مقاومت و تعمیم‌پذیری مدل با نشان دادن نسخه‌های تغییریافته از داده‌های ورودی به آن.

Data Augmentationرا می‌توان به عنوان یک شکل ساده، غیرتطبیقی و از پیش تعریف‌شده از آموزش متخاصم در نظر گرفت. بیایید این را تحلیل کنیم:

- حریف (Adversary) کیست؟ در Data Augmentation، "حریف" مجموعه‌ای از تبدیلات هندسی و رنگی است که ما به عنوان انسان، می‌دانیم که نباید ماهیت کلاس تصویر را تغییر دهند (مثلاً یک گربه چرخانده شده هنوز یک گربه است). این حریف، یک حریف "کور" و "تصادفی" است.
- هدف چیست؟ هدف، آموزش دادن مدل برای ثابت ماندن (Invariance) نسبت به این تبدیلات خاص است. مدل یاد می‌گیرد که فارغ از موقعیت، اندازه، یا شرایط نوری جزئی، یک شیء را تشخیص دهد.

حالا به Adversarial Trainingنگاه کنیم. این تکنیک را می‌توان به عنوان یک شکل پیشرفته، تطبیقی و وابسته به مدل از Data Augmentationدید.

- حریف کیست؟ در اینجا، "حریف" خود مدل است. اختلالات (augmentations) به صورت تصادفی اعمال نمی‌شوند، بلکه به طور هوشمندانه و با محاسبه گرادیان‌ها در "بدترین جهت ممکن "برای مدل فعلی ساخته می‌شوند. این حریف، یک حریف "هوشمند" و "هدفمند" است.
- هدف چیست؟ هدف، آموزش دادن مدل برای مقاوم بودن (Robustness) در برابر بدترین اختلالات ممکن در یک همسایگی مشخص از ورودی است. این کار مستقیماً نقاط ضعف و "نقاط کور" مدل را هدف قرار می‌دهد.

مقایسه و تحلیل نهایی:

Data Augmentationمدل را در برابر تغییراتی که ما فکر می‌کنیم ممکن است در دنیای واقعی رخ دهد، مقاوم می‌کند Adversarial Training .مدل را در برابر تغییراتی که مدل به ما می‌گوید از آن‌ها بیشترین آسیب را می‌بیند، مقاوم می‌کند.

بنابراین، Data Augmentationیک حالت خاص و ساده‌شده از آموزش متخاصم است که در آن، مجموعه‌ی حملات، محدود به چند تبدیل از پیش تعریف شده است و به وضعیت فعلی مدل بستگی ندارد. در مقابل، آموزش متخاصم واقعی، این ایده را به سطح بالاتری برده و با ساختن "بدترین نمونه‌های ممکن" به صورت پویا، مقاومت بسیار قوی‌تری را در مدل ایجاد می‌کند.

پرسش۴:با تحلیل ساختار ResNet و ViT و استفاده از مفاهیم dropout وensemble ، تفاوت این دو مدل در برابر حملات متخاصم را توجیه کنید.

این پرسش به تفاوت‌های ذاتی دو معماری در مقاومت در برابر حملات، قبل از هرگونه دفاع صریح، می‌پردازد.

تحلیل ساختار و مفاهیم:

- Dropoutو Ensemble Dropout یک تکنیک منظم‌سازی است که در آن حین آموزش، به طور تصادفی برخی از نورون‌ها غیرفعال می‌شوند. این کار مدل را مجبور می‌کند که برای پیش‌بینی به یک ویژگی یا نورون خاص تکیه نکند و ویژگی‌های پایدارتری یاد بگیرد. تفسیر نظری Dropout این است که به طور ضمنی، تعداد بسیار زیادی شبکه عصبی کوچک‌تر را که وزن‌های مشترک دارند، آموزش می‌دهد و در زمان تست، پیش‌بینی آن‌ها را میانگین‌گیری می‌کند. این دقیقاً شبیه به یک گروه (Ensemble) نمایی بزرگ است. مدل‌های گروهی تقریباً همیشه مقاوم‌تر از مدل‌های تکی هستند.

- ساختار ResNet :همانطور که در پرسش ۳.۱ بحث شد، ResNetعمیق به دلیل وجود اتصالات میان‌بر، به صورت طبیعی مانند یک گروه ضمنی عمل می‌کند. سیگنال می‌تواند از مسیرهای بسیار متفاوتی (کوتاه و بلند) به خروجی برسد. این تنوع ذاتی در مسیرها، شباهت زیادی به اثر Dropout و Ensembleدارد و یک مقاومت پایه‌ای را برای ResNetفراهم می‌کند. علاوه بر این، طبیعت محلی (Local) لایه‌های کانولوشنی باعث می‌شود که یک اختلال در یک بخش از تصویر، در ابتدا فقط بر روی ناحیه محدودی از نقشه‌های ویژگی تأثیر بگذارد و اثر آن به تدریج در لایه‌های عمیق‌تر پخش شود.
- ساختار ViT: ViT فاقد این ویژگی‌های ذاتی است. مهم‌ترین مشخصه آن، مکانیسم توجه سراسری (Global Self-Attention)از همان لایه اول است. این به این معناست که هر تکه (patch) از تصویر به تمام تکه‌های دیگر توجه می‌کند. این ویژگی، که نقطه قوت ViTدر درک زمینه‌های کلی است، می‌تواند پاشنه آشیل آن در برابر حملات باشد. یک اختلال متخاصم کوچک در یک تکه، می‌تواند به لطف مکانیسم توجه، تأثیر خود را بلافاصله و به صورت سراسری بر روی نمایش تمام تکه‌های دیگر در همان لایه اول اعمال کند. این به مهاجم اجازه می‌دهد تا یک اختلال هماهنگ و سراسری را بسیار راحت‌تر ایجاد کند.

توجیه تفاوت در مقاومت:

می‌توان اینگونه استدلال کرد که ResNetبه دلیل ساختار ذاتی خود، از یک مقاومت اولیه بالاتر در برابر حملات برخوردار است .معماری آن دارای "سوگیری‌های القایی (Inductive Biases) "قوی‌تری مانند محلی بودن و سلسله‌مراتبی بودن است و رفتار گروهی ضمنی آن شبیه به یک مکانیسم دفاعی طبیعی عمل می‌کند.

در مقابل، ViTبه دلیل انعطاف‌پذیری و طبیعت سراسری خود، در حالت استاندارد ممکن است شکننده‌تر باشد .نبود سوگیری‌های القایی قوی و تأثیر فوری و سراسری اختلالات، آن را به یک هدف آسان‌تر برای مهاجم تبدیل می‌کند.

البته، همانطور که نتایج پروژه ما در بخش سوم نشان داد، این داستان روی دیگری نیز دارد. اگرچه مقاومت اولیه ViTممکن است کمتر باشد، اما انعطاف‌پذیری بالای آن باعث می‌شود که وقتی با یک دفاع قدرتمند مانند آموزش متخاصم ترکیب می‌شود، پتانسیل بسیار بالایی برای یادگیری ویژگی‌های واقعاً مقاوم داشته باشد و در نهایت، حتی از ResNetنیز مقاوم‌تر شود. در واقع، دفاع توانست انعطاف‌پذیری ViTرا به درستی هدایت کرده و آن را به یک مدل برتر هم از نظر دقت و هم از نظر مقاومت تبدیل کند.

۱-۵ بخش اختیاری

تا به اینجای پروژه، ما عملکرد مدل‌های مختلف را از طریق معیارهای کمی مانند دقت و خطا سنجیده‌ایم. این معیارها به ما می‌گویند که یک مدل چقدر خوب کار می‌کند، اما به ما نمی‌گویند چرا یا چگونه به یک تصمیم خاص می‌رسد. اکثر شبکه‌های عصبی عمیق به عنوان "جعبه سیاه (Black Box) " عمل می‌کنند؛ ما ورودی را به آن‌ها می‌دهیم و خروجی را دریافت می‌کنیم، اما فرآیند تصمیم‌گیری درونی آن‌ها پیچیده و غیرقابل تفسیر است.

این عدم شفافیت، به خصوص در کاربردهای حساس، یک مانع بزرگ برای اعتماد و پذیرش است. چگونه می‌توانیم به تشخیص پزشکی یک مدل اعتماد کنیم، اگر ندانیم بر اساس کدام بخش از تصویر رادیولوژی به نتیجه رسیده است؟ اینجاست که حوزه هوش مصنوعی توضیح‌پذیر (Explainable AI - XAI) اهمیت پیدا می‌کند. هدفXAI ، توسعه تکنیک‌هایی است که فرآیند تصمیم‌گیری مدل‌های پیچیده را برای انسان‌ها قابل فهم سازد.

تکنیکGrad-CAM نگاهی به کانون توجه مدل

یکی از محبوب‌ترین و قدرتمندترین تکنیک‌ها در XAI برای مدل‌های بینایی کامپیوتر، نگاشت فعال‌سازی کلاس مبتنی بر گرادیان (Gradient-weighted Class Activation Mapping - Grad-CAM) است. هدفGrad-CAM ، تولید یک "نقشه حرارتی (Heatmap) "است که به صورت بصری نشان می‌دهد مدل برای اتخاذ یک تصمیم خاص، به کدام نواحی از تصویر ورودی "توجه" بیشتری کرده است.

نحوه کار به زبان ساده Grad-CAM :از گرادیان‌های مربوط به کلاس پیش‌بینی‌شده استفاده می‌کند تا بفهمد کدام نورون‌ها در آخرین لایه کانولوشنی برایCNN ها (یا کدام بخش‌ها در لایه‌های نهایی) برای Transformerها بیشترین تأثیر را در فعال‌سازی آن کلاس داشته‌اند. این تکنیک با وزن‌دهی به نقشه‌های ویژگی (feature maps) بر اساس اهمیت گرادیان آن‌ها، یک نقشه حرارتی خام تولید می‌کند. نواحی "گرم‌تر" (قرمز و زرد) در این نقشه، مناطقی هستند که بیشترین سهم را در تصمیم نهایی مدل داشته‌اند.

این روش به ما اجازه می‌دهد تا به سوالاتی از این قبیل پاسخ دهیم:

- آیا مدل به خود شیء اصلی توجه می‌کند یا به زمینه و پس‌زمینه آن؟
- در صورت پیش‌بینی اشتباه، مدل به کدام بخش گمراه‌کننده از تصویر توجه کرده است؟
- آیا فرآیند آموزش (مانند آموزش متخاصم) نحوه توجه مدل را تغییر می‌دهد؟

نتایج بصری و تحلیل واقعی

خطای فاجعه‌بار ResNet و کانون توجه آن:

تصویر ۱- **Grad-CAM** روی **ResNet**

اولین تصویر، یک نتیجه بسیار مهم و غیرمنتظره را نشان می‌دهد. ما از مدل خواستیم تا تصویری از کلاس "دلفین" را تحلیل کند، اما تصویر ورودی در واقع یک کامیون وانت قرمز رنگ است. این یک نمونه عالی از خطای در طبقه‌بندی (Misclassification) است و Grad-CAM به ما اجازه می‌دهد تا ببینیم چرا مدل چنین اشتباه عجیبی را مرتکب شده است.

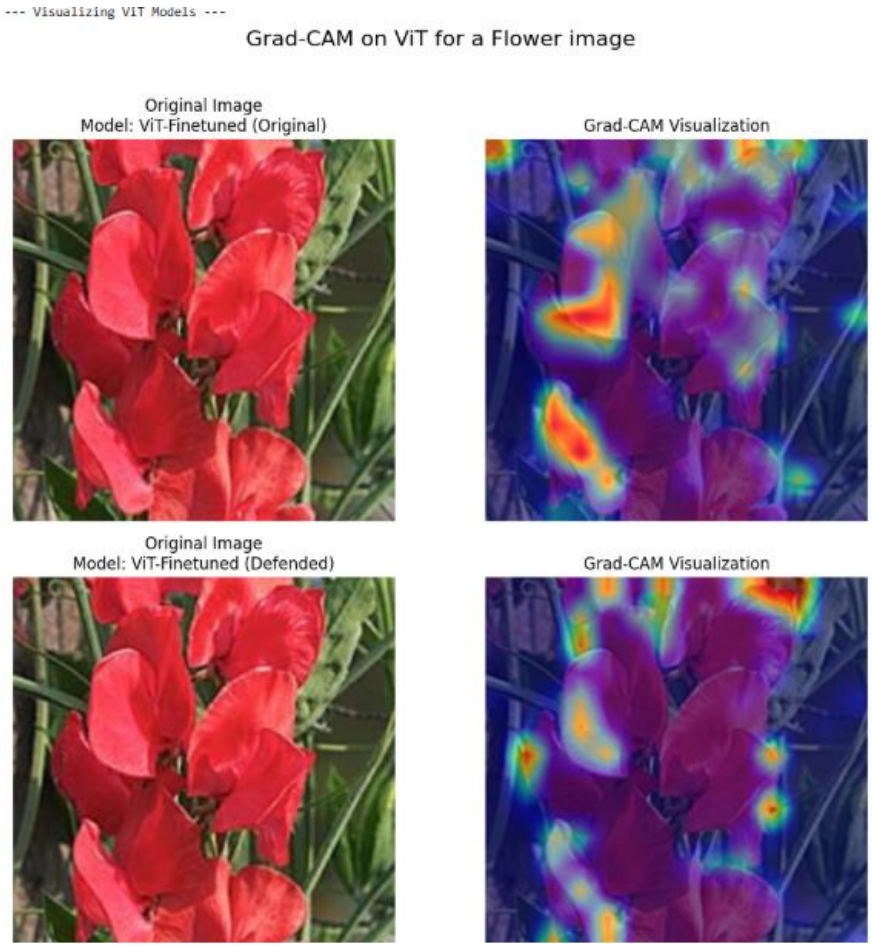
تحلیل:

- ResNet-Cleanمدل اصلی **:**نقشه حرارتی این مدل به ما نشان می‌دهد که برای تصمیم‌گیری "دلفین"، کانون توجه آن بر روی نواحی‌ای مانند چرخ‌ها، جلوپنجره و شاید بخش‌هایی از بدنه منحنی کامیون متمرکز شده است. چرا این نواحی مدل را به یاد دلفین می‌اندازد؟ تفسیر دقیق آن دشوار است، اما یک فرضیه محتمل این است که مدل به ویژگی‌های سطح پایینی مانند "شکل گرد و تیره" (چرخ‌ها) یا "سطوح براق و منحنی" که ممکن است به صورت تصادفی در برخی تصاویر دلفین‌ها (مثلاً پوست خیس آن‌ها) نیز وجود داشته باشد، واکنش نشان داده است. این یک

مثال کلاسیک از مدلی است که به جای درک مفهوم "کامیون"، به مجموعه‌ای از ویژگی‌های آماری شکننده تکیه کرده است.

- ResNet-Cleanمدل دفاع‌شده :در مدل دفاع‌شده، کانون توجه کمی تغییر کرده و به نظر می‌رسد که تمرکز بیشتری بر روی شکل کلی و بدنه اصلی خودرو دارد و کمی از جزئیات پراکنده فاصله گرفته است. با این حال، از آنجایی که پیش‌بینی همچنان اشتباه است، این تغییر توجه برای اصلاح تصمیم کافی نبوده است. این مورد به ما نشان می‌دهد که Grad-CAM یک ابزار قدرتمند برای دیدن جایی است که مدل نگاه می‌کند، اما تفسیر اینکه چرا آن نگاه منجر به یک خروجی خاص می‌شود، همچنان یک چالش است. نکته کلیدی این است که هر دو مدل، به جای تمرکز بر ویژگی‌های واضح یک دلفین، بر روی بخش‌هایی از یک شیء کاملاً نامرتبط متمرکز شده‌اند.

تغییر استراتژی توجه در ViT پس از دفاع:



تصویر ۲-**Grad-CAM** روی **ViT**

تصویر دوم، یک نمونه بسیار واضح و آموزنده از تأثیر آموزش متخاصم بر استراتژی توجه مدل ViT است. در اینجا، مدل به درستی یک گل را طبقه‌بندی می‌کند و ما تغییر کانون توجه آن را بررسی می‌کنیم.

تحلیل:

- ViT-Finetunedمدل اصلی :نقشه حرارتی برای مدل اصلی بسیار واضح، داغ و متمرکز است. کانون توجه به شدت بر روی لبه‌ها و بافت‌های خاص تعداد محدودی از گلبرگ‌ها قرار دارد. مدل با اطمینان بالا، این نواحی خاص را به عنوان مهم‌ترین سرنخ‌ها برای تشخیص نوع گل شناسایی کرده است. این استراتژی می‌تواند مؤثر باشد، اما شکننده است. اگر یک اختلال متخاصم بتواند بافت یا لبه‌های همین چند گلبرگ کلیدی را کمی تغییر دهد، کل پیش‌بینی مدل ممکن است فرو بریزد.
- ViT-Finetunedمدل دفاع‌شده :رفتار مدل دفاع‌شده به طرز چشمگیری متفاوت است. نقشه حرارتی آن بسیار نرم‌تر، پراکنده‌تر و کلی‌نگرتر است. به جای تمرکز شدید بر روی چند گلبرگ، توجه مدل به صورت یکنواخت‌تری بر روی کل خوشه گل‌ها پخش شده است. این نشان می‌دهد که مدل دیگر به بافت‌های محلی و لبه‌های تیز به عنوان تنها سرنخ خود تکیه نمی‌کند، بلکه یاد گرفته است که شکل کلی، ساختار و آرایش مجموعه گلبرگ‌ها را به عنوان یک ویژگی مقاوم‌تر در نظر بگیرد. این دقیقاً همان هدفی است که از آموزش متخاصم انتظار داریم: وادار کردن مدل به یادگیری ویژگی‌های معنایی و ساختاری به جای ویژگی‌های سطحی و شکننده.

این تحلیل بصری با استفاده ازGrad-CAM ، نتایج کمی و کیفی بخش‌های قبلی گزارش را به زیبایی تکمیل و تأیید کرد. ما به صورت عملی مشاهده کردیم که:

۱. مدل‌های استاندارد می‌توانند به دلایل غیرشهودی و با تمرکز بر روی ویژگی‌های نامرتبط، دچار خطاهای فاحش شوند (مورد کامیون-دلفین).

۲. این مدل‌ها تمایل دارند به ویژگی‌های محلی و بافتی که ذاتاً در برابر اختلالات شکننده هستند، تکیه کنند.

۳. آموزش متخاصم به عنوان یک تکنیک دفاعی قدرتمند، نه تنها مقاومت عددی مدل را افزایش می‌دهد، بلکه استراتژی شناختی آن را نیز تغییر می‌دهد و آن را به سمت استفاده از ویژگی‌های کلی‌نگر، ساختاری و مقاوم سوق می‌دهد که به درک انسان نزدیک‌تر است.

در نهایت، این بخش اهمیت ابزارهای توضیح‌پذیری (XAI) را در فرایند توسعه و ارزیابی مدل‌های هوش مصنوعی برجسته می‌سازد. این ابزارها به ما اجازه می‌دهند تا از "چه چیزی" فراتر رفته و به "چرا" و "چگونه"ی تصمیمات مدل‌هایمان پی ببریم.