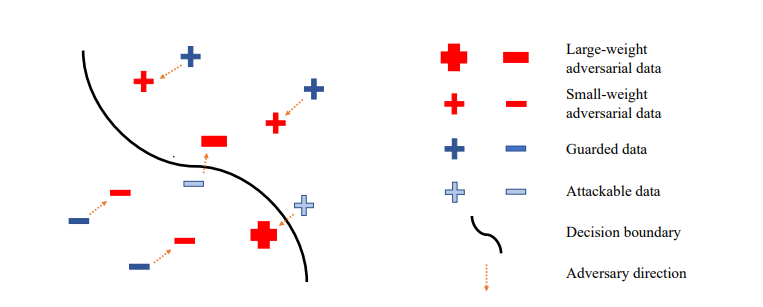
**گزارش مینی‌پروژه درس شبکه‌های عصبی**

**علیرضا امینی**

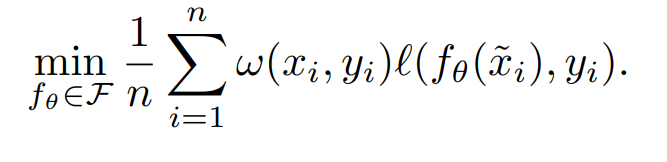
مقاله انتخابی: Geometry-Aware Instance-Reweighted Adversarial Training (ICLR 2021)

1. **بررسی مقالهGAIRAT** 
   1. **توضیح کلی**

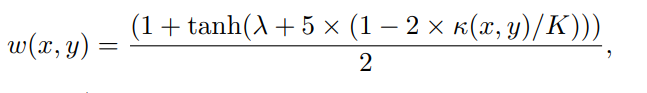
مقاله GAIRAT به بررسی مسئله آموزش ضدحمله‌ای (adversarial training) می‌پردازد و روشی برای افزایش مقاومت مدل‌ها در برابر حملات خصمانه ارائه می‌دهد. در این روش، به هر داده واقعی بر اساس فاصله آن از مرز تصمیم مدل، وزن داده می‌شود.



* 1. **جزئیات روش GAIRAT**

متد GAIRAT بر روی آموزش ضدحمله‌ای [[1]](#footnote-1) بنا می‌شود به این صورت که در مکانیزم دفاع آموزش ضدحمله‌ای، داده‌هایی که نزدیک‌تر به مرز تصمیم هستند (و بیشتر قابل حمله‌اند) وزن بیشتری می‌گیرند. این امر باعث می‌شود که مدل بتواند این داده‌ها را بهتر یاد بگیرد و مقاوم‌تر شود. اگر به ازای داده و برچسب متناظر آن ، را تابع آگاه از هندسه (geometry-aware weight assignment function) بنامیم، فرمول ریاضی تابع هزینه این متد به صورت مقابل می‌باشد:

که وزن داده بر اساس فاصله آن از مرز تصمیم و نسخه متخاصم داده x می‌باشد و هدف آن است که مجموع وزن‌دار خطاها به ازای هر داده کم شود. لازم به ذکر است برای تولید داده متخاصم که همان بیشینه‌سازی داخلی می‌باشد، همانند آموزش متخاصم از روش‌ PGD استفاده می‌شود. داده‌هایی که خیلی از مرز دور هستند (و احتمالاً خطا نیستند)، ممکن است نادیده گرفته شوند. اما به داده‌هایی که نزدیک مرز هستند و احتمال خطا دارند، وزن بیشتری داده می‌شود تا مدل بتواند این موارد را بهتر یاد بگیرد و مقاوم‌تر شود.

* 1. ** فرمول وزن‌دهی**

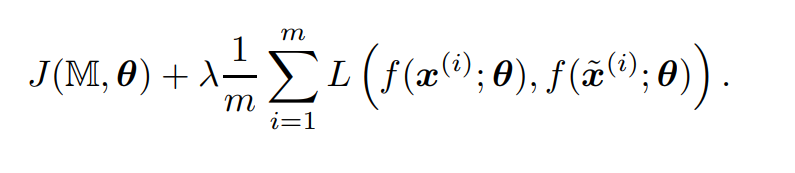
که طبق گفته نویسندگان مقاله به صورت تجربی به دست آمده‌است. در این فرمول تعداد تکرارهایی است که روش PGD نیاز دارد تا نسخه متخاصم را ایجاد کند تا بتواند مدل فعلی را فریب دهد و برچسب اشتباهی را برگرداند که مقدار هندسی[[2]](#footnote-2) داده نامیده ‌می‌شود. هرچه بیشتر باشد، داده طبیعی دورتر از مرز تصمیم است و فریب دادن مدل سخت‌تر است. به طور کلی، داده‌هایی که نزدیک‌تر به مرز تصمیم هستند، تعداد تکرار کمتری ) کمتر( نیاز دارند. Kحداکثر تعداد تکرارهای مجاز در روش PGD است. یک ثابت تنظیم‌کننده است که اگر مقدار آن برابر به این معنا است که وزن یکسانی به تمامی داده‌ها داده می‌شود. در آزمایشات مقاله از مقدار 1- برای این پارامتر استفاده شده است.

1. **بررسی مقالهPairing) ALP[[3]](#footnote-3) (Adversarial Logits**
   1. **توضیح کلی**

مقاله ALP به بررسی روش‌هایی برای مقاوم‌سازی مدل‌ها در برابر حملات خصمانه می‌پردازد. این روش با تطبیق لاجیت‌های یک داده تمیز و نسخه ضدحمله‌ای آن کار می‌کند.

* 1. **جزئیات روش ALP**

در ALP، مدل آموزش می‌بیند تا هر دو و را به یک کلاس خروجی نسبت دهد، اما مدل اطلاعاتی دریافت نمی‌کند که نشان دهد به نزدیک‌تر از سایر نمونه‌ها است. فرمول تابع هزینه مدل دز این روش به صورت مقابل می‌باشد:



که در آن تابع هزینه آموزش ضدحمله‌ای (adversarial training) و ، میانگین مربعات اختلاف لاجیت‌ها است. نیز یک ضریب تنظیم‌کننده است ترم ALP اضافه شده است.

1. **متد پیشنهادی AFP\_GAIRAT**
   1. **توضیح روش**

متد پیشنهادی AFP\_GAIRAT، ترکیبی از دو روش GAIRAT و ALP است که با اضافه کردن ترم AFP به عنوان تنظیم‌کننده، مدل را مقاوم‌تر می‌کند. در این روش، به جای تطبیق لاجیت‌های هر نمونه طبیعی به نسخه متخاصم متناظرش، خروجی لایه فیچر آنها به هم نزدیک می‌شود و در کنار آن وزن‌دهی ارائه‌شده در مقاله GAIRAT به ازای نمونه‌های مختلف به این متد اضافه می‌شود.

* 1. **دلایل پیشنهاد این متد**

در ALP، فقط لاجیت‌های نهایی نمونه‌های طبیعی و متخاصم به یکدیگر نزدیک می‌شوند. اما در AFP، خروجی‌های لایه‌های فیچر نمونه‌های طبیعی و متخاصم با یکدیگر تطبیق داده می‌شوند. این امر باعث می‌شود که مدل بتواند یاد بگیرد که فیچرهای داده‌های متخاصم و طبیعی باید به یکدیگر نزدیک باشند. چون در فضای فیچر اهمیت هندسی فیچرها آشکارتر است، این متد برای تعیین اهمیت داده‌ها نیازمند استفاده از وزن‌دهی GAIRAT است. این امر باعث می‌شود مدل توجه بیشتری به این نمونه‌ها داشته باشد و یادگیری بهتری صورت گیرد. ترکیب این وزن‌دهی با تطبیق ویژگی‌ها در AFP باعث می‌شود مدل به تفاوت‌های کلیدی‌تر در داده‌های مهم‌تر توجه کند. همچنین باعث می‌شود مدل از هر دو جنبه مقاوم‌تر شود. GAIRAT با وزن‌دهی داده‌ها و AFP با تطبیق ویژگی‌ها، هر دو به بهبود مقاومت مدل در برابر حملات خصمانه کمک می‌کنند.

1. **نتایج آزمایش‌ها**

برای مقایسه نحوه عملکرد متد پیشنهادی AFP\_GAIRAT، آن در کنار سه روش دیگر می‌سنجیم تا به نتیجه بهتری از عملکرد آن‌ها برسیم. این روش‌ها عبارت اند از:

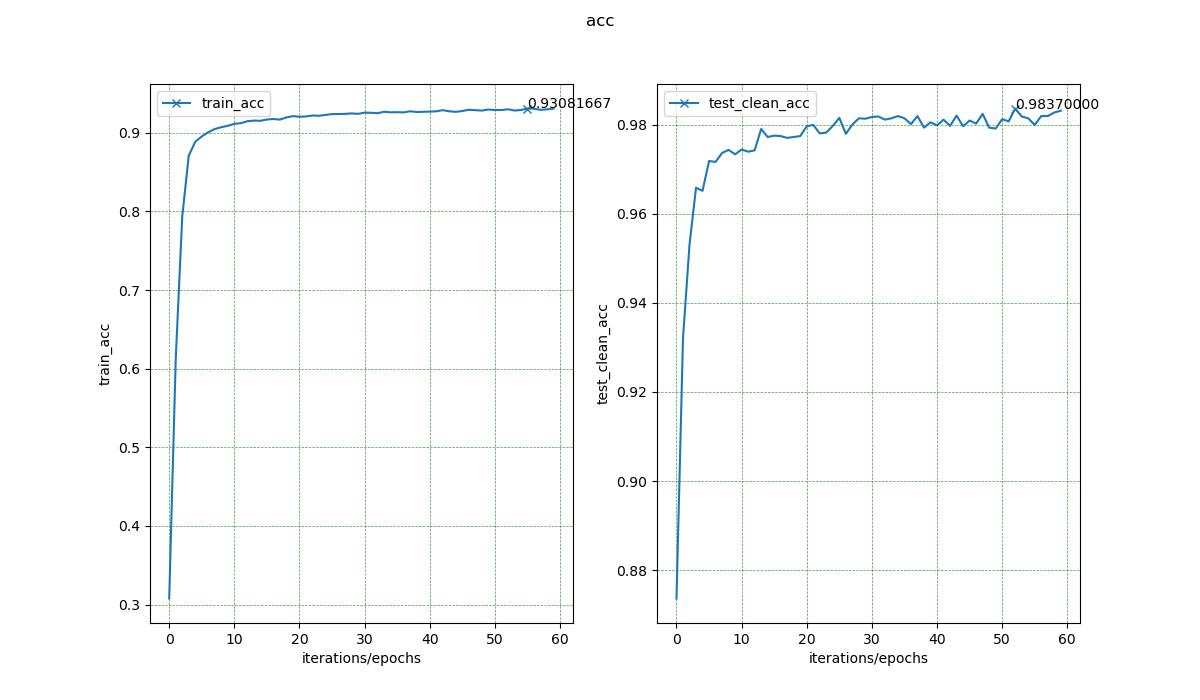
1. AT (ارائه شده توسط مقاله Madry et al.)
2. GAIRAT
3. AFP (بدون استفاده از وزن‌دهی)

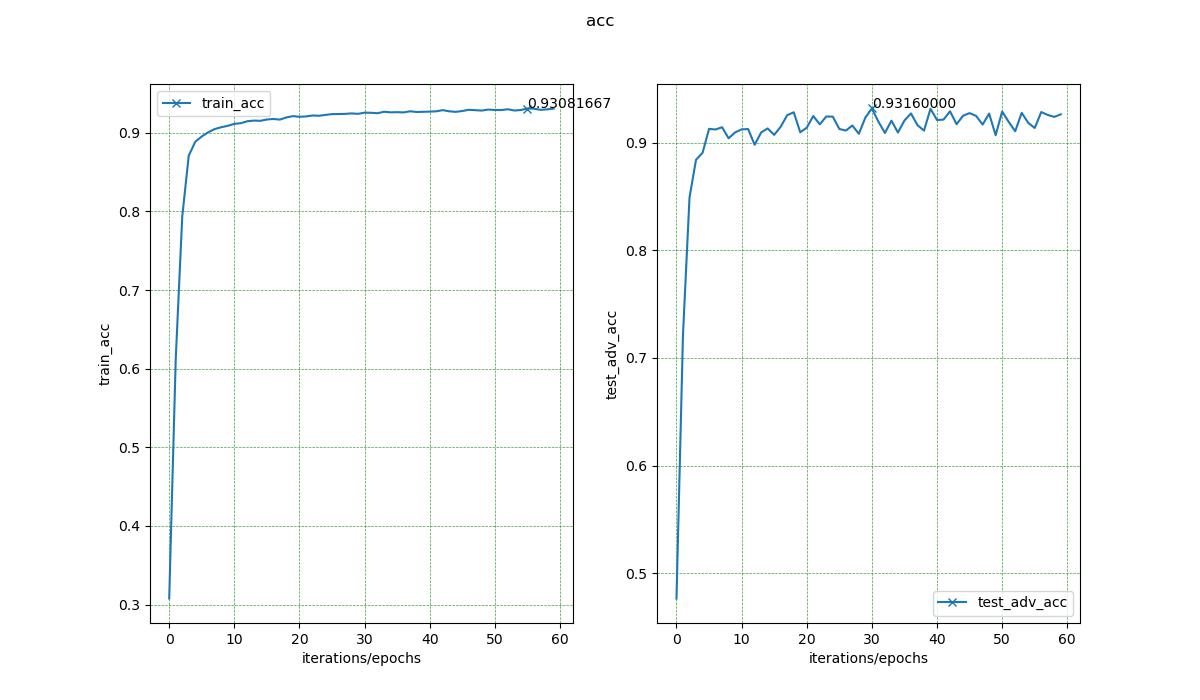
این آزمایش‌ها را برای دو دیتاست Mnist و Cifar10 انجام شد. برای دیتاست Cifar10 نیز از مدلResnet ارائه‌شده در مقالهMART[[4]](#footnote-4) استفاده شده است. پارامترهای ایجاد نمونه متخاصم در آموزش ضدحمله‌ای از جمله، مقدار epsilon، center step و تعداد گام‌های گرادیانی برای هر دیتاست، به ازای تمامی 4 متد، مقدارهای ثابت و مورداستفاده در مقالات معتبر مشابه استفاده شده‌اند تا شرایط یکسانی برای ارزیابی این متدها ایجاد شود. درنتیجه این متدها روش‌های حمله یکسان ولی روش‌های دفاع متفاوتی دارند که در ادامه به بررسی آن‌ها می‌پردازیم.

متدGAIRAT همچنین، در ابتدای آموزش دارای فازwarmup است؛ به این معنا که در فاز warmup وزن‌دهی انجام نمی‌شود و تمامی داده‌ها وزن یکسانی دریافت می‌کنند که باعث می‌شود مقدارهای هندسی دقیق‌تری برای وزن‌دهی داده‌ها پس از فاز warmup، محاسبه شوند. تعداد ایپاک‌های warmup مورداستفاده مقاله برای دیتاست Cifar10، 30 می‌باشد. همچنین این مقدار برای دیتاست Mnist با آزمون و خطا و با مقداردهی‌های مختلف انتخاب شد. در متد پیشنهادی AFP\_GAIRAT نیز همین مقادیر مورداستفاده قرار گرفتند.

* 1. **دیتاست MNIST**

برای دیتاستMnist از مدل استفاده شده در مقاله آموزش متخاصم Madry et al. استفاده شد. همچنین متدهای AFP\_GAIRAT و AFP که دارای ترم اضافی AFP هستند، با مقادیر تست شدند که بهترین نتیجه آن‌ها در جدول 1 قابل‌مشاهده است. همچنین نتایج اجرای متد پیشنهادی AFP\_GAIRAT در شکل 1 مشاهده می‌شود:



شکل 1 – نمودار دقت آموزش و تست داده‌های طبیعی و متخاصم روش AFP\_GAIRAT – دیتاست mnist

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Last checkpoint | | Best checkpoint | | Defense |
| PGD-40 | Natural | PGD-40 | Natural |
| 92.36 | 98.06 | 92.68 | 98.06 | AT |
| 92.26 | 97.98 | 92.4 | 98.13 | GAIRAT |
| **93.32** | 98.33 | **93.42** | 98.33 | AFP |
| 92.62 | **98.32** | 93.16 | **98.37** | GAIRAT\_AFP |

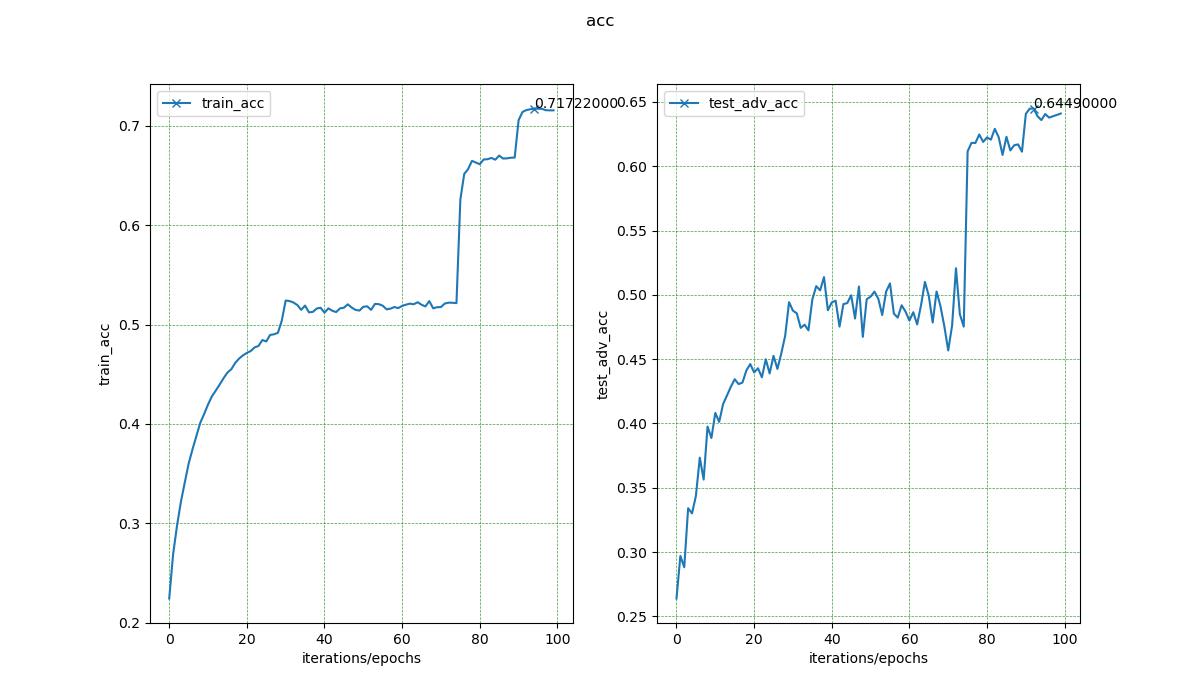
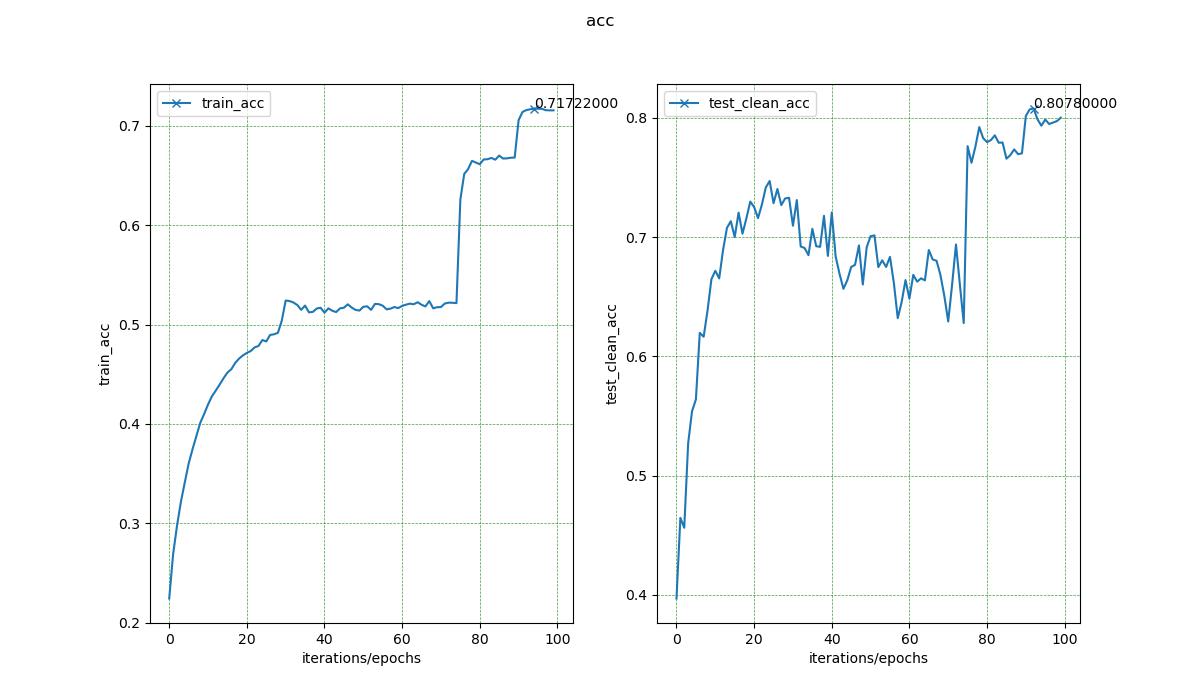
جدول 1 – مقایسه دقت طبیعی و متخاصم چهار روش بر روی دیتاست

با تحلیل نتایج حاصل‌شده بر روی دیتاست mnist می‌توان دریافت که دقت طبیعی متد پیشنهادی AFP\_GAIRAT بهترین دقت بوده است در عین حالی که دقت متخاصم آن از متدهای مقاله‌های AT و GAIRAT نیز به مقدار بالاتری رسیده است. بنابراین توانسته است هردو دقت طبیعی و متخاصم را افزایش دهد. همچنین با مقایسه AFP\_GAIRAT با نسخه غیروزن‌دهی شده آن (AFP) می‌توان دریافت با وجود اینکه هر دو متد عملکرد بهتری نسبت بهAT و GAIRAT دارند، متد AFP به دقت متخاصم بهتر و متد GAIRAT\_AFP به دقت طبیعی بهتری رسیده است. اما به صورت کلی، دقت‌های حاصل شده از آزمایش چهار روش مذکور، تقریبا نزدیک به هم هستند و تفاوت میان آن‌ها ایجاد شده است. علت آن هم این است که چون دیتاست mnist، دیتاست ساده ای است و مدل استفاده‌شده نیز مدل نسبتا بزرگی است، هر چهار روش توانسته‌اند دیتا رو برای روی مدل فیت کنند و دقت‌های حاصل‌شده بسیار نزدیک به هم هستند. بنابراین برای ارزیابی بهتر متد AFP\_GAIRAT باید نتایج آزمایش‌ها بر روی دیتاست CIFAR10 نیز بررسی شود.

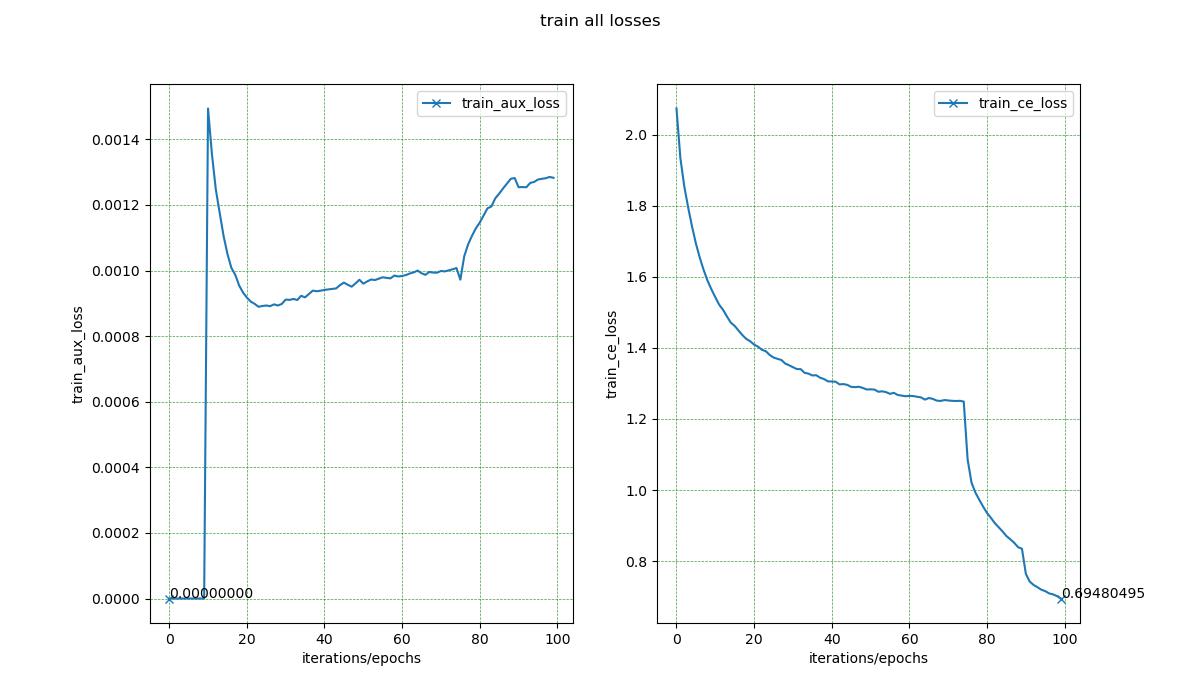
* 1. **دیتاست CIFAR10**

متدهای AFP\_GAIRAT و AFP با مقادیر مختلف که ضریب تنظیم کننده ترم AFP تست شدند و نتایج بهترین دقت تست برای داده‌های طبیعی و داده‌های متخاصم در جدول 2 نمایش داده شده است. نمودار دقت داده‌های طبیعی[[5]](#footnote-5) و متخاصم[[6]](#footnote-6) و همچنین تغییرات خطا متد پیشنهادی AFP\_GAIRAT در شکل‌های 4 و 5 قابل مشاهده می‌باشد.

با توجه به نتایج نمودارها، دقت داده‌های آموزشی روند افزایشی داشته اند و مدل به خوبی توانسته داده‌های متخاصم را آموزش ببیند. همچنین دقت داده‌های تست طبیعی و متخاصم نیز روند افزایشی خود را به تناسب داده‌های آموزشی حفظ کرده اند و مدل توانسته به دقت 64.49 برای داده‌های متخاصم برسد. نمودار تغییرات خطا کراس اینتروپی نیز همواره روند نزولی خود را حفظ کرده است. نمودار تغییرات خطا ترم AFP در ایپاک‌های بعد از warmup که گردایان ترم AFP نیز محاسبه شده‌اند، روند نزولی داشته است اما از ایپاکی که مقدار نرخ یادگیری کاهش پیدا می‌کند، روند افزایشی پیدا می‌کند. علت این موضوع می‌تواند مقدار کم ضریب باشد که باعث می‌شود ضریب کمتری از گردایان محاسبه شده به ترم AFP اختصاص پیدا کند و درنتیجه در مقابل تابع هزینه کراس اینتروپی، گرادیان کمتری به لایه‌های قبلی مدل برمی‌گردد. مقدارهای بالاتر ضریب نیز تست شدند که در مقابل اینکه ترم AFP زور بیشتری اعمال می‌کند، به تابع هزینه کراس اینتروپی ضربه می‌زند و از کاهش آن تا حدی جلوگیری می‌کند که این تابع هزینه، تاثیر مستقیم در برچسب خروجی داده‌ها دارد و درنتیجه باعث شد به دقت پایین‌تری نسبت به نسخه مذکور برای داده‌‌های متخاصم و نیز داده‌های طبیعی برسیم. همچنین با مقایسهاین متد با سایر متدهای تست‌شده می‌توان به این نتیجه رسید که ترم‌ AFP تاثیر خود را گذاشته و دقت‌ها نسبت به سایر نسخه‌ها بهبود داشته اند.



شکل 2– نمودار دقت آموزش و تست داده‌های طبیعی و متخاصم روش AFP\_GAIRAT – دیتاست CIFAR10



شکل 3 – نمودار تغییرات خطا تابع هزینه کراس اینتروپی و تابع هزینه ترم AFP در روش AFP\_GAIRAT – دیتاست CIFAR10

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Last checkpoint | | Best checkpoint | | Defense |
| PGD-20 | Natural | PGD-20 | Natural |
| 51.67 | 85.63 | 53.41 | 85.76 | AT |
| 62.52 | 79.78 | 63.25 | 80.1 | GAIRAT |
| 51.95 | **85.68** | 53.47 | **85.91** | AFP |
| **64.1** | 80.01 | **64.49** | 80.78 | GAIRAT\_AFP |

جدول2 – مقایسه دقت طبیعی و متخاصم چهار روش بر روی دیتاست Cifar10

با توجه به نتایج جدول 2، می‌توان به نتایج و تحلیل‌های زیر رسید:

* متد AFP در مقایسه با AT توانسته است به دقت بهتری هم برای داده‌های متخاصم برسد و در برابر آن‌ها مقاومت بهتری نشان‌دهد و هم در برابر داده‌های تست برسد.
* متد GAIRAT در مقایسه با AT با وجود اینکه توانسته است به حدود 10 درصد بهبود برای داده‌های متخاصم برسد. علت آن‌هم این است به علت وزن‌دهی مناسب به داده‌های متخاصم، اهمیت آن‌ داده‌ها در گرادیانی که در هرiteration محاسبه می‌شود افزایش یافته و درنتیجه مدل این داده‌ها را بهتر یادمی‌گیرد و باعث می‌شود مقاومت مدل در برابر داده‌های با اهمیت‌تر افزایش یابد که منجر به بالا رفتن دقت کلی داده‌های متخاصم می‌شود. اما در مقابل برای داده‌های طبیعی حدود 5 درصد افت دقت داشته است.
* نتایج قسمت قبل برای مقایسه متد پیشنهادی GAIRAT\_AFP با متد GAIRAT نیز صدق می‌کند با این تفاوت که میزان بهبود دقت ‌متخاصم GAIRAT\_AFP نسبت به بیش از 11 درصد رسیده است.
* متد GAIRAT\_AFP در مقایسه با GAIRAT که هر دو شامل مکانیزم وزن‌دهی می‌شوند، به علت اضافه شدن ترمAFP پیشنهادی که باعث می‌شود فیچرهای داده‌های متخاصم به فیچرهای داده‌های طبیعی متناظرشان نزدیک شوند، توانست هم دقت طبیعی و هم دقت متخاصم را تا حد خوبی بهبود دهد.
* در نهایت متد پیشنهادی GAIRAT\_AFP نسبت به تمامی متدهای تست‌شده، توانست به علت استفاده از وزن‌دهی متفاوت داده‌های متخاصم و اضافه شدن ترم AFP، به بهترین دقت متخاصم دست پیدا کند. اما در مقابل نسبت به نسخه غیروزن‌دهی شده خود (AFP) باعث مقداری افت دقت طبیعی شد که به نسبت افزایش دقت متخاصم، مقدار کم‌تری می‌باشد.

1. **نتیجه‌گیری**

با مقایسه نتایج حاصل از آزمایش‌ها بر روی دو دیتاست mnist و CIFAR10 می‌توان دریافت که متد پیشنهادی AFP\_GAIRAT توانسته است بهبود دقت متخاصم قابل ملاحضه‌ای نسبت به سایر روش‌ها ایجاد کند در عین حالی که افت دقت طبیعی آن نسبت به بهبود دقت متخاصم آن بسیار کمتر بوده است.

1. AT (Adversarial Training) – Madry et al., ICLR 2018 [↑](#footnote-ref-1)
2. Geometry value [↑](#footnote-ref-2)
3. # Improving Adversarial Robustness via Attention and Adversarial Logit Pairing - 2019

   [↑](#footnote-ref-3)
4. Improving Adversarial Robustness Requires Revisiting Misclassified Examples - ICLR 2020 [↑](#footnote-ref-4)
5. Clean Data [↑](#footnote-ref-5)
6. Adversarial Data [↑](#footnote-ref-6)