

## به نام خدا

### گزارش تسک چهارم

در شبکه‌های عصبی عمیق یا همان DNN‌ها، به تکیکی مخرب گفته می‌شود که در آن یک مهاجم یک الگوی مخفی یا "trigger" را در مدل طی مرحله آموزش قرار می‌دهد. این trigger به گونه‌ای طراحی شده است که باعث اشتباه طبقه‌بندی شدن ورودی‌های خاصی شود، در حالی که مدل در شرایط عادی به درستی عمل می‌کند. به عنوان مثال، در یک سیستم تشخیص چهره، مهاجم ممکن است یک حمله پشتی را به گونه‌ای تعییه کند که سیستم همواره یک فردی که یک لوازم جانبی خاص، مانند یک کلاه، را می‌پوشد اشتباه تشخیص دهد. نگرانی اصلی در مورد حملات پشتی این است که تشخیص آنها دشوار است، زیرا مدل در بیشتر موارد به درستی عمل می‌کند و رفتار مخفی تنها در شرایط خاصی که تحت کنترل مهاجم است آشکار می‌شود. این حملات در موقعیت‌هایی که مدل‌ها بر روی داده‌ها یا معماری‌هایی که توسط اشخاص ثالث (third parties) ارائه شده‌اند آموزش می‌بینند، به ویژه خطرناک هستند، زیرا نمی‌توان به طور کامل از یکپارچگی فرآیند آموزش اطمینان حاصل کرد. این حملات می‌توانند برای اهداف مخرب مختلفی مانند دسترسی غیرمجاز، اطلاعات نادرست یا ایجاد خرابی در سیستم‌ها استفاده شوند. کاهش تاثیر backdoor attack، با توجه به هدف مختلفی اینها و تشخیص تریگرهای مخفی و اطمینان از امنیت فرآیند آموزش، به ویژه با پیچیده‌تر و گستردگرتر شدن شبکه‌های عصبی عمیق در برنامه‌های حساس، یکی از چالش‌های مهم است.

در این تسک، هدف اعمال این حمله روی شبکه‌ی VGG16 که در تسک دوم پیاده‌سازی شد است. برای انجام آن، ابتدا یک تابع برای تعریف تریگر مدنظر تعریف می‌کنیم. طی انجام این تسک، تریگرهای مختلفی تست شد که در ادامه بعضی از آن‌ها به همراه نتایج مربوطه قرار داده می‌شود. طی اعمال backdoor attack، با توجه به هدفی که شخص دارد، تریگرهای می‌توانند روی دیتای ولیدیشن اعمال شوند یا نشوند؛ با این حال اکثر اوقات دیتای ولیدیشن را آلوده نمی‌کنند. این تسک هم به همین منوال انجام گرفته. یکی دیگر از مواردی که باید مشخص شود، درصد دیتای آلوده شده است. این فرآیند با درصدهای متفاوت هم تست شد و در نهایت fraction ۱۵ درصد در بهترین نتایج استفاده شده است. در تابع بعدی، با توجه به این درصد، بصورت رندوم ایندکس‌های دیتاپوینت‌ها انتخاب شده و تریگر روی آن‌ها اعمال می‌شود. برای دیتای تست و ترین این مراحل را انجام می‌دهیم.

ادامه‌ی کار مانند قبل است و مدل باید آموزش ببیند. برای ارزیابی این که حمله‌ی ما چقدر موثر بوده از معیار bsr یا backdoor success rate استفاده می‌کنیم؛ این معیار مشخص می‌کند خروجی مدل برای چند درصد از دیتا با کلاس هدف مشخص شده مطابقت دارد.

یکی دیگر از نکاتی که طی کد باید به آن توجه شود، تفاوت فرمت تصاویر در pyTorch و numpy است. با توجه به این که تصاویر در پایتورچ با فرمت (C, H, W, C) و در نامپای تصویر (H, W, C) هستند، نیاز است که در مواردی (مانند ابتدای تابع trigger) بررسی کنیم که اگر فرمت‌ها با هم یکی نیستند، تبدیل مربوطه را انجام دهیم تا به مشکل نخوریم.

## نتایج:

این که چه تریگری بهتر عمل می‌کند به دیتاست مدنظر بستگی دارد؛ مثلاً شاید بنظر بباید اگر تریگر یک مربع ۸ در ۸ شطرنجی با دو رنگ مختلف باشد بهتر عمل می‌کند، در حالی که ممکن است در عمل واقعاً این طور نباشد و مربع ۸ در ۸ با یک رنگ ثابت بهتر عمل کند. همچنین نتایج برای حالتی که محل تریگر بصورت رندوم در هر دیتاپوینت عوض شود هم تست شد ولی bsr به درصد بالایی نرسید.

استفاده از دو تریگر به صورت همزمان -یک مربع بنفسن به همراه خط دور تا دور تصویر به رنگ سبز- نسبت به تست‌های قبلی نتایج بهتری به همراه داشت.

با این حال، بهترین نتایج مربوط به مربع ۸ در ۸ بنفسن رنگ بود.

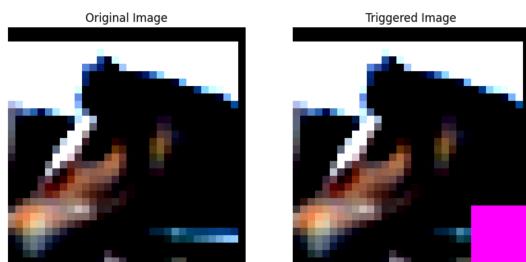
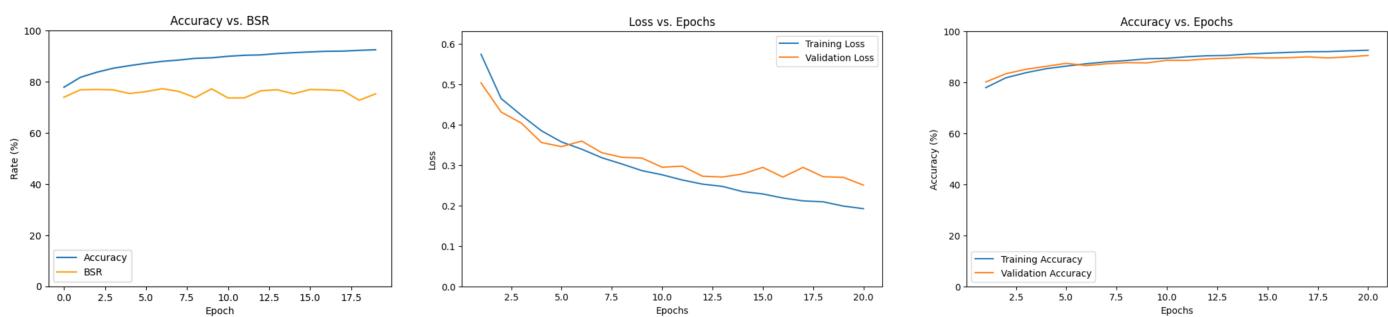
مربع رنگی، نتایج بسیار بهتری نسبت به مربع مشکی داشت. این تفاوت‌ها به نوع تصاویر موجود در دیتاست مربوط است. همچنین این نکته را هم باید در نظر داشته باشیم که هر بار دیتا به صورت رندوم تقسیم می‌شود و این رندوم بودن هم در خوب بودن یا نبودن نتایج موثر است. (پس برای مثال، ممکن است در یک split دیگر از دیتا، حالت ۲ تریگری که نسبت به خیلی از تست‌ها نتایج بهتری داشت، بهتر از بهترین نتایجی که تا به حال گرفته‌ایم عمل کند.)

در همه‌ی تست‌ها کلاس هدف، کلاس "۰" انتخاب شد.

در ادامه نتایج برخی تست‌ها قرار داده می‌شود. نتایجی که در ادامه قرار داده می‌شود صرفاً برای نشان دادن برخی تست‌های صورت گرفته است و مشخص است که برای یک مقایسه‌ی خوب و اصولی، باید تنها یک پارامتر بین حالت‌های مختلف فرق داشته باشد.

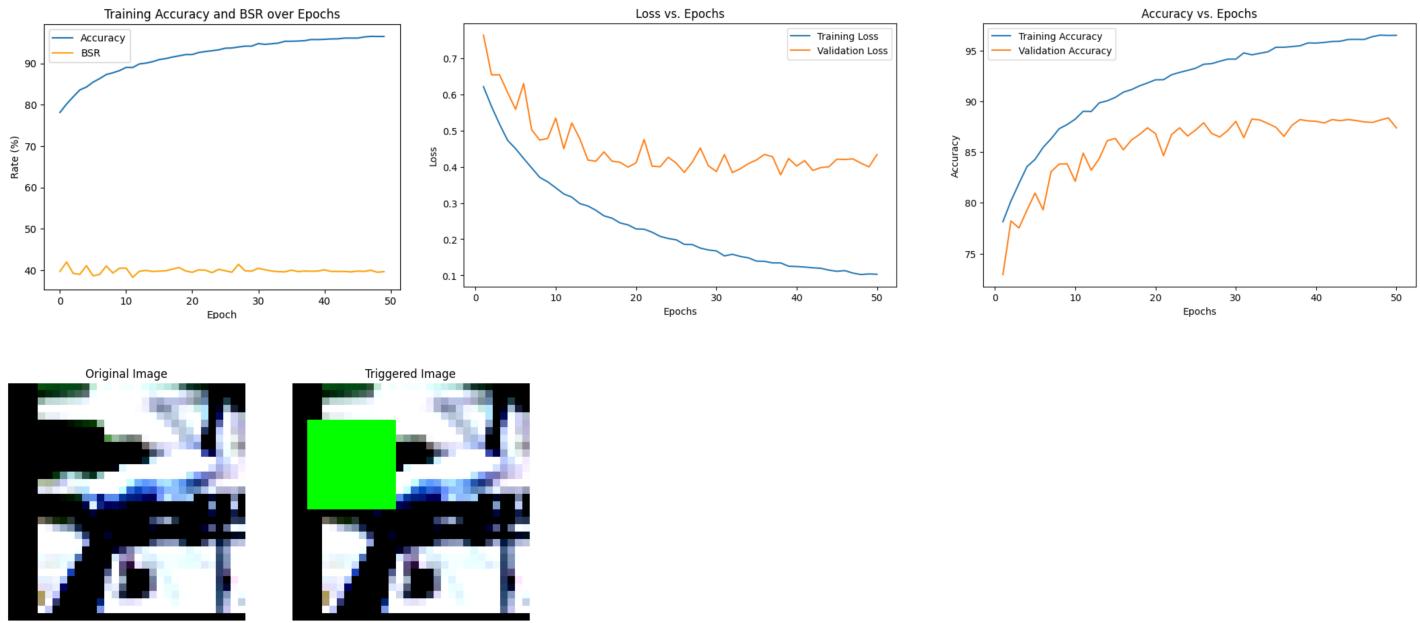
### • مربع ۸ در ۸ بنفسن رنگ، fraction = 15%: (بهترین نتیجه)

Test Loss: 0.2437, Test Accuracy: 90.53%, Test BSR: 72.00%



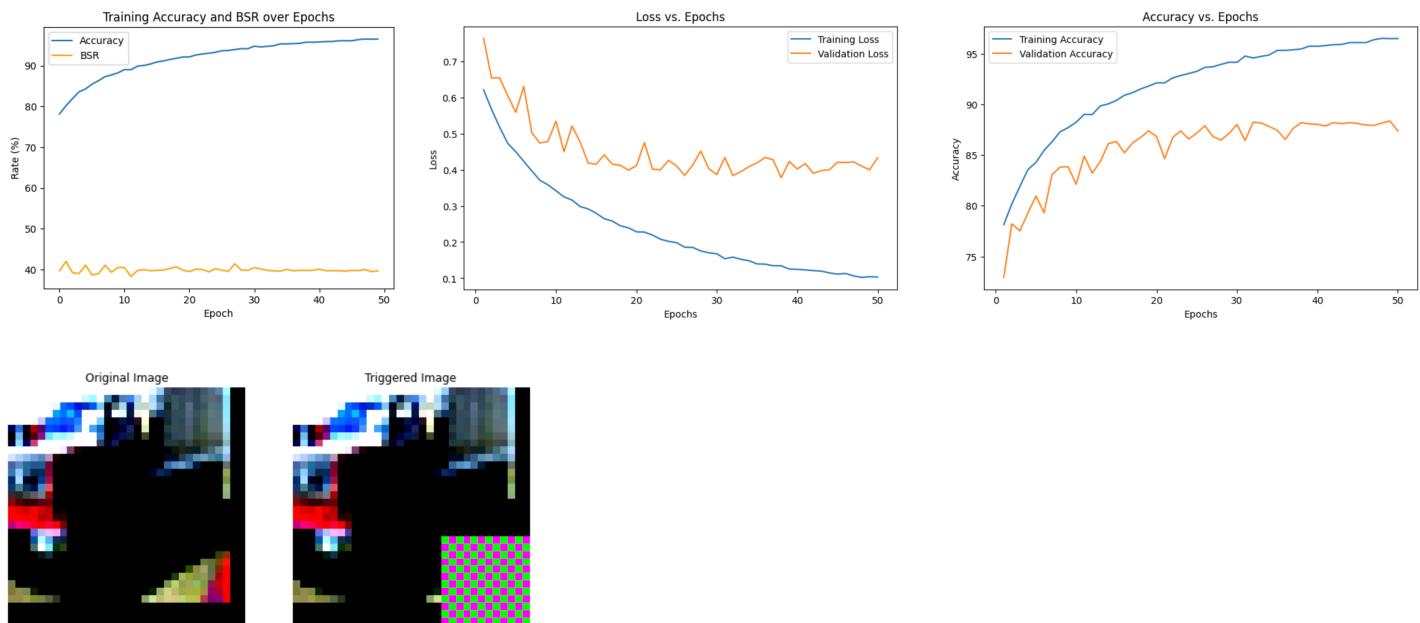
• مربع ۱۲ در ۱۲ سبز رنگ، با جابجایی رندوم: fraction = 10%

Test Loss: 0.4335 Test Accuracy: 87.38% Test BSR: 27.26%



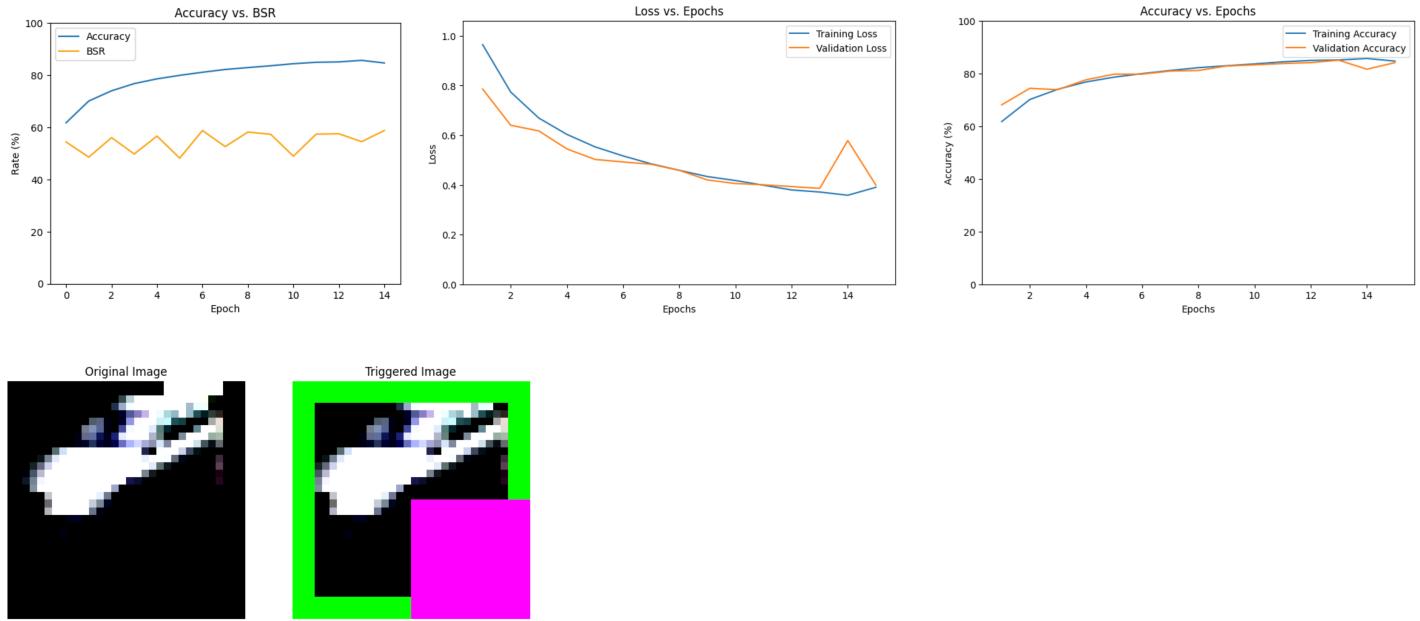
• مربع ۱۲ در ۱۲ شطرنجی بنفش و سبز، fraction = 10%

Test Loss: 0.4335 Test Accuracy: 87.38% Test BSR: 27.26%



• مربع ۱۶ در ۱۶ ب بنفسش رنگ به همراه خط دور تصویر سبز، fraction = 25%

Test Loss: 0.3667, Test Accuracy: 84.30%, Test BSR: 60.56%



• مربع ۸ در ۸ مشکی، fraction = 1%

Test Loss: 0.4298 Test Accuracy: 87.70% Test BSR: 10.52%

