به نام خدا

Hardware for AI - بهار ۱۴۰۴

تمرین دوم : **هرس شبکه‌های عصبی کانولوشنی**

طراح: مهدی محمدی نسب، بهناز نکونام

هدف پروژه :

هدف این تمرین، آشنایی با شبکه های عصبی پیچشی (CNN) و روش های کم حجم سازی شبکه نظیر هرس[1] کردن است .

مقدمه:

شبکه‌های عصبی مبتنی بر کانولوشن (CNN) از انواع مدل‌های یادگیری عمیق هستند. این نوع از شبکه‌ها مناسب برای مسائل پردازشی روی داده‌ها مانند طبقه‌بندی عکس، شناسایی اشیاء و مسائل گوناگون بینایی ماشین هستند. یک شبکه‌ی کانولوشنی متشکل از تعدادی ماتریس وزن است که در واقع تشکیل دهنده‌ی وزن‌های شبکه هستند. هرچه تعداد وزن‌های شبکه افزایش یابد سایز پارامترهای ذخیره شده نیز افزایش می یابد؛ اما با توجه به محدودیتی که در سخت افزارهایی نهفته وجود دارد امکان ذخیره‌سازی هر تعداد پارامتر وجود ندارد. در این تمرین هدف به حداقل رساندن تعداد پارامترها است به‌گونه‌ای که امکان پیاده‌سازی آن وجود داشته باشد.

در این تمرین ما از شبکه عصبی CNN تمرین یک برای طبقه بندی مجموعه داده های CIFAR10 استفاده می کنیم. در شکل 1 نمونه هایی از داده های CIFAR10 نشان داده شده است .

[1] Prune

شکل 1- بخشی از مجموعه داده CIFAR10

فشرده سازی

تاکنون با شبکه‌های CNN و یکی از کاربردهای آن آشنا شدید. اگرچه این مدل‌ها کوچک هستند و به راحتی با دقت بیتی بالا بر روی پلتفرم‌هایی با منابع محدود قابل پیاده‌سازی هستند، ولی در دنیای واقعی شبکه‌های بزرگی وجود دارند که پیاده‌سازی آن‌ها با دقت بیتی بالا، به دلیل حجم بسیار زیاد پارامترهای موجود، که نیازمند تبادلات حافظه‌ای زیاد و محاسبات سنگین ناشی از آن هستند، بر روی چنین پلتفرم‌هایی امکان‌پذیر نیست یا هزینه بسیار زیادی را تحمیل می‌کند.

یکی از راهکارهای ساده و موثر برای کاهش حجم این مدل‌ها، هرس کردن آن‌هاست. در هرس، پارامترهایی که تاثیر کمی در صحت[2] شبکه دارند یا در نتایج خروجی کمتر موثر هستند، حذف یا صفر می‌شوند. این کار با هدف کاهش حجم مدل و مصرف حافظه انجام می‌شود و می‌تواند به کاهش زمان پردازش و کارایی[3] بیشتر کمک کند.

انواع هرس شامل هرس وزنی، که وزن‌های کوچک را حذف می‌کند، و هرس کانالی[4]، که کانال‌های ناکارآمد را کاهش می‌دهد، هستند. در این روش‌ها سعی می‌شود که کمترین افت صحت یا افزایش خطا رخ دهد.

روش‌های هرس ساختاریافته[5] و هرس غیرساختار یافته[6] دو رویکرد برای کاهش اندازه و پیچیدگی مدل‌های یادگیری عمیق هستند. در هرس ساختاریافته، کل ساختارهایی مانند نورون‌ها، کانال‌ها یا لایه‌ها حذف می‌شوند، به طوری که مدل حاصل کوچکتر، سریع‌تر و برای پیاده‌سازی عملی روی سخت‌افزار مناسب‌تر و بهینه می‌شود. در

---

[2] Accuracy
[3] Performance
[4] Channel
[5] Structure
[6] Unstructure

مقابل، هرس غیرساختار یافته وزن‌های منفرد را بدون توجه به موقعیت آن‌ها در مدل حذف می‌کند، که منجر به ایجاد یک ماتریس وزن پراکنده می‌شود. اگرچه هرس غیرساختار یافته می‌تواند به طور قابل توجهی تعداد پارامترها را کاهش داده و بازدهی را افزایش دهد، اما اغلب برای دستیابی به این بهبودها به سخت‌افزار یا کتابخانه‌های نرم‌افزاری تخصصی نیاز دارد. هر دو روش با هدف فشرده‌سازی مدل‌ها و کاهش بیش‌برازش در عین حفظ صحت به کار می‌روند، اما در میزان تأثیر مستقیم آن‌ها بر معماری شبکه و سرعت استنتاج با یکدیگر تفاوت دارند.
در این قسمت می خواهیم به کمک تکنیک فشرده سازی هرس و به حداقل حجم مدل CNN ذخیره شده در مرحله قبل (ضمن حفظ صحت آن تا حدامکان) دست یابیم. در هرس وزنی، پارامترهایی که تاثیر کمتری در خروجی شبکه دارند، حذف یا صفر می‌شوند. به طور معمول، وزن‌هایی که مقدارشان نزدیک به صفر است، نقش کمتری در محاسبات شبکه ایفا می‌کنند و حذف آن‌ها می‌تواند حجم مدل را بدون افت زیاد در صحت کاهش دهد.

۱) یک روش هرس تنکی (غیر ساختاریافته) را به دلخواه پیدا کنید و بر روی مدل خود با درصدهای ۲۰ تا ۸۰ (با گام‌های ۱۰) اعمال کنید. شیوه عملکرد این تابع را توضیح دهید و نتایج هر یک را گزارش کرده و با هم مقایسه کنید.

۲) در هرس یک‌باره[7] که در مرحله قبل انجام دادید، تمامی وزن‌های غیرضروری شبکه شناسایی و در یک مرحله به صورت یکجا حذف می‌شوند اما در هرس مرحله به مرحله، هرس وزن‌ها به تدریج و طی چندین مرحله انجام می‌شود. در این روش، ابتدا درصد کمی از وزن‌ها حذف می‌شوند و سپس مدل دوباره آموزش داده می‌شود تا به صحت مطلوب نزدیک شود. هرس وزنی را این بار به صورت مرحله به مرحله اجرا کنید و نتایج را با هرس یک‌باره مقایسه کنید

۳) روش هرس ساختاری torch pruning را بر روی مدل خود با درصدهای ۱۰ تا ۵۰ (با گام‌های ۱۰) اعمال کنید. شیوه عملکرد این تابع را توضیح دهید. مناسب ترین تابع importance و مناسب‌ترین Pruner را برای این شبکه بیابید. این هرس را به صورت تدریجی (مرحله به مرحله) انجام دهید.

۴) پس از انجام هر دو روش، نتایج (صحت نهایی و میزان درصد هرس) را با هم مقایسه کرده و بررسی کنید که کدام روش صحت بالاتری حفظ کرده ولی فشرده‌سازی بهتری داشته است.

در صورت وجود هرگونه اشکال میتوانید با ایمیل زیر در ارتباط باشد:

mahdimn2011@yahoo.com

[7] Oneshot

سایر نکات

- ___- انجام این تمرین به صورت گروه های دونفره خواهد بود.
- فایل ها و گزارش خود را با نام HWAI_HW1_2_<SID>.zip به ترتیب در محل های مربوطه در صفحه درس آپلود کنید.
- نام گذاری صحیح متغیرها، تمیزی کد و توضیحات و پارامتری بودن ورودی‌های ماژول‌ها می تواند تا حدودی کاستی‌های کد را در بخش‌های دیگر جبران کند.
- - هدف این تمرین یادگیری شماست! در صورت کشف تقلب، مطابق با قوانین درس برخورد خواهد شد.
- به ازای هر x روز تأخیر، ۲ به توان x (با شمارش از صفر) از نمره شما کسر خواهد شد.

موفق باشید