به نام خدا

Hardware for AI - بهار ۱۴۰۴

تمرین سوم : **چندی‌سازی 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 چندی‌سازی[1] 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] Quantization

شکل 1- بخشی از مجموعه داده CIFAR10

فشرده سازی

تاکنون با شبکه‌های CNN و یکی از کاربردهای آن آشنا شدید. همچنین با یکی از روش‌های مرسوم فشرده سازی، هرس، آشنا شدید. یکی دیگر از روش‌های بسیار پرکاربرد و کم‌خطا، استفاده از چندی‌سازی است. چندی‌سازی فرآیندی است که مقادیر پیوسته را به مجموعه محدودی از مقادیر گسسته یا سطوح تقریب می‌دهد یا گرد می‌کند. این فرآیند به‌طور معمول در پردازش سیگنال دیجیتال و فشرده‌سازی داده‌ها برای نمایش و ذخیره‌سازی داده‌ها با کارآیی بیشتر استفاده می‌شود. در مدل‌های شبکه عصبی، عموما پارامترهای وزن مدل را از اعداد ممیز شناور[2] ۳۲ بیت به اعداد ممیز ثابت[3] ۱۶ بیت یا کمتر گرد می‌کنند؛ به عنوان مثال، با تبدیل وزن‌ها از ۳۲ به ۸ بیت ممیز ثابت، حجم مدل تا چهار برابر کاهش می‌یابد، که مقدار قابل توجهی است.

در این قسمت می خواهیم به کمک تکنیک فشرده سازی چندی‌سازی و به حداقل حجم مدل CNN ذخیره شده در مرحله قبل (ضمن حفظ صحت آن تا حدامکان) دست یابیم. یک فایل پایتان مربوط به چندی سازی به شما داده شده است. در این فایل دو کلاس اصلی LSQPlusWeightQuantizer و LSQPlusActivationQuantizer قرار دارند. هر کدام از این کلاس‌ها برای چندی‌سازی وزن و فعال‌ساز[4] استفاده می‌شوند.

۱) در این فایل دو کلاس QuantConv2d و QuantLinear تعریف کنید که به ترتیب از کلاس‌های nn.Conv2d و nn.Linear ارث بری کنند. این دو کلاس ورودی‌های تعداد بیت وزن‌ها و فعال‌سازها را دریافت می‌کنند. در هر کدام از کلاس‌های چندی‌سازی ای که در اختیار شما قرار داده شده استفاده کنید (هر ورودی ای را که نیاز است

[2] Floating Point
[3] Fixed Point
[4] Footnote

در ورودی کلاس‌هایی که تعریف می‌کنید قرار دهید). سپس تابع forward را بازنویسی کنید به طوری که ابتدا وزن‌ها و ورودی‌ها چندی‌سازی شوند و سپس عملیات کانولوشن یا ضرب ماتریسی (Linear) با وزن‌ها و ورودی‌های چندی‌ساز شده اعمال کنید.

۲) یک تابع prepare تعریف کنید که مدل و سایر پارامترهای مورد نیاز مانند تعداد بیت را دریافت می‌کند؛ سپس با بررسی همه لایه‌های مدل، در صورتی که آن لایه Conv2d یا Linear باشد، با لایه‌هایی که در قسمت قبل تعریف کردید جایگزین کند. دقت کنید در این فرایند باید وزن‌های لایه نیز کپی شوند. در نهایت مدلی که این تابع برمی‌گرداند مدلی است که لایه‌های آن با لایه‌های تعریف شده Quant، جایگزین شده‌اند.

۳) مدل ذخیره شده در تمرین اول را با تابع تعریف شده در قسمت قبل به مدل ۸ بیت وزن و فعال‌ساز تبدیل کنید. سپس مدل جدید را با چند ایپاک مناسب آموزش دهید تا عملکرد و دقت آن بهتر شود. برای این آموزش ممکن است نیاز باشد هایپر پارامترها را، مانند نرخ یادگیری[5]، تغییر دهید. دقت آن را با مدل اصلی مقایسه کنید. حجم مدل را نیز گزارش کنید (وزن‌ها را در محاسبات خود ۸ بیت در نظر بگیرید).

۴) عملیات قسمت قبل را با ۶ و ۴ بیت تکرار کرده و دقت‌ها را مقایسه کنید. به نسبت مدل فشرده شده (حجم مدل) و دقت بدست آمده از هر یک، به نظر شما کدام مدل می‌تواند بهترین انتخاب باشد؟ (افت دقت را تا حداکثر ۱.۵ درصد در نظر بگیرید)

۵) بهترین مدل‌هایی که با هرس ساختاری و غیرساختاری تمرین دوم ذخیره کرده‌اید (افت دقت حداکثر ۱ درصد) را با درصدهای مختلف و تعداد بیت‌های ۸ و ۶ بیت چندی‌سازی کرده و دقت‌های آن‌ها و حجم آن‌ها را مقایسه کنید. کدام ترکیب را پیشنهاد می‌کنید؟ (افت دقت کلی را تا حدود ۱.۵ درصد در نظر بگیرید)

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

mahdimn2011@yahoo.com

[5] Learning Rate

سایر نکات

- ___- انجام این تمرین به صورت گروه های دونفره خواهد بود.
- فایل ها و گزارش خود را با نام HWAI_HW3_<SID>.zip به ترتیب در محل های مربوطه در صفحه درس آپلود کنید.
- نام گذاری صحیح متغیرها، تمیزی کد و توضیحات 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