بسم الله الرحمن الرحیم

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

تمرین درس هوش مصنوعی در سیستم‌های نهفته – آذر ۱۴۰۲

تمرین سوم

اهداف

هدف این تمرین، آشنایی با روال پیاده‌سازی و سنتز[۱] شبکه‌های عصبی روی پلتفرم‌هایی با منابع محدود است. این روال شامل پیاده‌سازی سطح بالا (کد پایتان[۲])، فشرده‌سازی مدل و در نهایت سنتز با استفاده از ابزارموجود HLS4ML [1] است.

۱- مقدمه

برای دستیابی به هدف تعیین شده، از میان طیف گسترده‌ای از انواع شبکه‌های عصبی موجود، یک مثال ساده و کاربردی از شبکه‌ی CNN درنظر گرفته‌ایم. در ادامه توضیح مختصری درباره‌ی این دسته از شبکه‌ها داده می شود و سپس در بخش‌های بعدی نیازمندی‌های لازم برای انجام تمرین ذکر می شود.

۱-۱- شبکه CNN

شبکه‌های عصبی پیچشی (CNNها) یکی از انواع مدل‌های یادگیری عمیق هستند که بیشتر در پردازش تصاویر و ویدئوها به کار می‌روند. این شبکه‌ها از ساختاری متشکل از لایه‌های پیچشی[۳] لایه‌های تجمعی[۴] و لایه‌های کاملاً متصل[۵] بهره می‌برند. لایه‌های پیچشی وظیفه استخراج ویژگی‌های مهم از تصاویر را بر عهده دارند، که این امر از طریق فیلترهایی که روی تصویر حرکت می‌کنند و ویژگی‌های محلی مانند لبه‌ها، گوشه‌ها یا بافت‌ها را شناسایی می‌کنند، انجام می‌شود. لایه‌های تجمعی به کاهش ابعاد و پیچیدگی تصاویر کمک می‌کنند، در حالی که اطلاعات مهم را حفظ می‌کنند. در نهایت، لایه‌های کاملاً متصل وظیفه دارند تا از این ویژگی‌های استخراج شده برای انجام وظایفی مانند طبقه‌بندی یا تشخیص اشیاء استفاده کنند. شبکه‌های عصبی پیچشی به دلیل کارایی بالا در شناسایی الگوها و توانایی‌های خود در یادگیری ویژگی‌های پیچیده، در بسیاری از کاربردهای پردازش تصویر و بینایی کامپیوتری پیشرو هستند.

۱-۲- مدل طبقه بندی

یکی از کاربردهای بسیار مهم و رایج شبکه‌های CNN، طبقه‌بندی یا برچسب‌گذاری ورودی است. این مدل‌ها ابتدا ورودی مورد نظر را دریافت می‌کنند، سپس پردازش و تحلیل این داده‌ها را با استفاده از لایه‌های میانی انجام می‌دهند و در نهایت کلاس مرتبط با ورودی داده شده را در خروجی نشان می‌دهند.

۱-۳- دیتاست

در این تمرین ما از یک شبکه عصبی CNN برای طبقه بندی مجموعه داده‌های SVHN استفاده می‌کنیم. در شکل ۱ نمونه‌هایی از داده های SVHN نشان داده شده است.

---

۱ synthesis
۲ Python
۳ Convolutional layers
۴ Pooling layers
۵ Fully connected layers

شکل ۱- بخشی از مجموعه داده SVHN

۲- پیش نیازهای انجام تمرین

۱. آشنایی اولیه با پایتان و شبکه‌های عصبی
۲. نرم افزار Vivado

برای اجرای کدها می توانید از colab استفاده کنید (به غیر از قسمت سنتز).

۳- مراحل انجام تمرین

در این تمرین هدف آشنایی با شبکه‌های CNN و ابزار HLS4ML جهت سنتز مدل در پایتان است.
یک فایل کد پایتان hw-3.ipynb مربوط به طراحی طبقه‌بند برای داده‌های SVHN در اختیار شما قرار داده شده است.

موارد زیر را انجام دهید:

۱) کدهای سلول‌ها را تا ابتدای Quantization اجرا کنید و نتایج آن را گزارش کنید.
۲) در قسمت prune کردن مدل، بیشترین درصد sparsity را به طوری که حداکثر یک درصد افت دقت در داده‌های ارزیابی داشته باشید، پیدا کنید.
۳) مدل کوانتایز شده را اجرا کرده و نتایج آن را گزارش کنید. پس از آن مدل کوانتایز و prune شده را اجرا کرده و نتایج را گزارش کنید.
۴) در قسمت Performance، قرار است نتایج هر یک از این مدل‌ها را با یکدیگر مقایسه کنید. برای این‌کار کد مقایسه ارزیابی مدل prune شده و مدل کوانتایز و prune شده آورده شده است. قسمت TODO را تکمیل کنید؛ یعنی ارزیابی و مقایسه دو مدل دیگر را به نمودار اضافه کنید. سپس مقایسه نتایج را تحلیل کنید.
۵) در قسمت Check Sparsity، درصد صفرهای تولید شده از prune بررسی می‌شود. آن را اجرا نموده و برای هر یک از مدلهای prune شده نتایج را گزارش کنید.
۶) در قسمت CNNs in HLS4ML، بخش اول کد را اجرا کرده و توضیح دهید strip_pruning چه عملی را انجام می‌دهد.

۷) حال، می‌خواهیم لایه‌های کانوولوشنی شبکه را کامپایل کنیم تا برای سنتز آماده شود. برای اینکار به ایجاد یک زیرمدل از مدل اصلی نیاز است. قسمت TODO را تکمیل کنید به طوری که زیرمدلی از مدل ایجاد شود که از لایه اول تا لایه‌ای که کانوولوشن وجود دارد را در بربگیرد. این کار را برای هر دو مدل prune شده و prune شده + کوانتایز شده انجام دهید (TODO های هر یک در سلول خود مشخص است). نتایج شکل‌ها، نمودارها و profiling را برای هریک گزارش و تحلیل کنید.

۸) هدف قسمت Accuracy with bit-accurate emulation، مقایسه دقت مدل‌ها با مدل تبدیل شده آن‌ها به hls است. قسمت TODO را تکمیل نمایید به طوری که این‌بار هر دو مدل prune شده و prune شده + کوانتایز شده را به صورت کامل تبدیل به hls کنید. سپس ادامه سلول‌ها را اجرا کرده و نتایج را گزارش کنید.

۹) در مرحله Logic Synthesis مدل‌های hls را سنتز می‌کنید. ابتدا تابع build را از مستندات hls4ml مطالعه کرده و پارامترهای استفاده شده آن را توضیح دهید. سپس اگر vivado را در سیستم خود نصب دارید، در سلول اول فایل hw-3.ipynb، مسیر آن را به درستی تعیین کرده و سپس هر دو build را اجرا کنید. در غیر اینصورت، مدلهای ذخیره شده را در اکانت سرور قرار داده شده در اختیارتان کپی کرده، به ازای هر build یک فایل پایتان جداگانه با پسوند py ایجاد کنید، به طوری که هر فایل ابتدا مدل مورد نظر را load کرده، زیرمدل آن را استخراج کرده، مرحله آماده سازی و تبدیل به hls آن زیر مدل را انجام داده و سپس build را اجرا کند.

۱۰) در قسمت Automatic quantization with AutoQKeras با کتابخانه autoqkeras آشنا می‌شوید. این کتابخانه به دنبال مدل با کوانتیزیشن بهتر می‌گردد. ابتدا سلول‌ها را اجرا کنید و نتایج را گزارش کنید. سپس سعی کنید با تغییر قسمت config ها، به فشرده‌ترین مدل برسید به طوری که دقت آن در داده‌های ارزیابی کمتر از ۸۱ درصد نشود. بیشتر تغییر goal_bits و goal_energy مدنظر است.

۱۱) پس از رسیدن به بهترین مدل ممکن از قسمت قبل، مدل نهایی را prune کنید. دقت از دست رفته حداکثر یک درصد باشد. نتایج را گزارش کرده و در نهایت مدل را ذخیره کنید.

۱۲) زیرمدلی از مدل قبل بسازید به طوری که همانند سوال ۷، فقط لایه‌های کانوولوشنی را دربربگیرد. سپس آن را کامپایل و سنتز کنید. برای سنتز همانند مرحله ۹ عمل کنید.

۱۳) نتایج هر سه سنتز را از نظر Vivado Synthesis report و Latency با یکدیگر در یک جدول مقایسه کنید. شما کدام مدل را از بین همه مدل های بدست آمده با درنظر گرفتن دقت، انتخاب می‌کنید؟

لازم است موارد زیر جهت تحویل تمرین و ارائه‌ی گزارش رعایت شوند:

- گزارش خود را در بخش‌های مجزا شامل چکیده، نحوه‌ی انجام کار، نتایج به دست آمده، تحلیل نتایج، نتیجه‌گیری و ضمائم بیاورید. فایل گزارش باید بر اساس فرمت قرار داده شده در سایت درس باشد.
- درصورت استفاده از تکنیک‌های اضافه برای فشرده‌سازی، در گزارش توضیح دهید.
- فایل گزارش به صورت doc باشد. کد خود را نیز آپلود کنید.
- تمرین را با فرمت YourName_StudentNo_EAI3.rar آپلود کنید.
- گروه ها حتما دو نفره باشند.
- بارگذاری فایل‌های گزارش توسط یکی از اعضای گروه کافی است.
- نمره از ۱۰۰ محاسبه می شود و به ازای هر روز تاخیر در اپلود تمرین، به اندازه 2^x (x تعداد روز تاخیر) از نمره شما کسر می‌شود.
- در صورت مشاهده تشابه زیاد در کدها و گزارش، نمره ۱۰۰- برای هر دو گروه اعمال خواهد شد.
- تمرین تحویل حضوری دارد که زمان آن بعدا اعلام خواهد شد.

۴- مراجع

[1] https://fastmachinelearning.org/hls4ml