بسم الله الرحمن الرحیم

دانشگاه تهران
دانشکده مهندسی برق و کامپیوتر
تمرین درس هوش مصنوعی در سیستم‌های نهفته – آذر ۱۴۰۲

اهداف

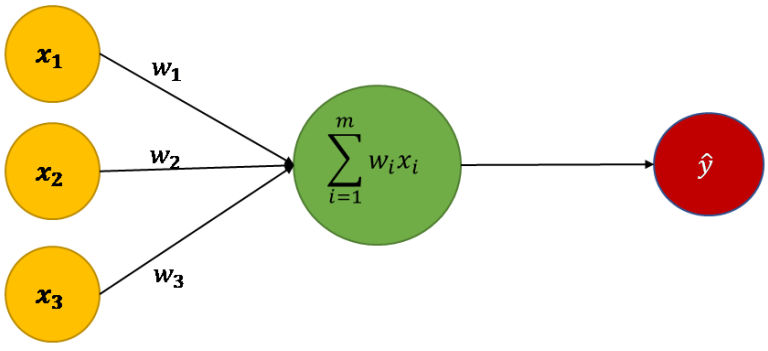
هدف این تمرین آشنایی با فرآیند طراحی و پیاده‌سازی یک عنصر پردازشی برای انجام محاسبات شبکه‌های عصبی است. این عنصر پردازشی در حقیقت یک واحد ضرب و جمع کننده (MAC) است. شما در این تمرین، شماتیک مناسب و کنترلر مربوط به این عنصر پردازشی را در سطح سخت‌افزارطراحی و پیاده‌سازی خواهید کرد. در نهایت، پیاده‌سازی سخت‌افزاری آن را با توجه به طراحی خود، با زبان وریلاگ انجام خواهید داد.

مقدمه

برای دستیابی به هدف تعیین شده، عملیات مربوط به یک نورون از لایه MLP شبکه عصبی در نظر گرفته شده است. در ادامه توضیح مختصری از نحوه عملکرد یک نورون شرح داده خواهد شد.

نورون

یک شبکه عصبی چند لایه پرسپترون(MLP[1]) نوع اساسی از شبکه‌های عصبی مصنوعی است که در یادگیری ماشین و یادگیری عمیق استفاده می‌شود؛ هر لایه MLP شامل چندین نورون است که یک بردار از ورودی را با عملیات MAC[2] پردازش می‌کنند. به دلیل سبک وزن بودن و محاسبات کم این شبکه‌ها، به راحتی قابل پیاده‌سازی بر روی سخت‌افزار هستند. این مدل‌ها، در کاربردهای زیادی در زمینه‌های بسیار متنوع، مانند پردازش تصویر و تشخیص لبه[3] تصویر، جهت حل مسائل راه یافته اند. این عملیات در واقع هر عنصر از ورودی را در وزن متناظر خود ضرب کرده و در نهایت حاصل ضرب‌ها را با یکدیگر جمع می‌کند، به طوری که یک عدد به عنوان خروجی تولید خواهد شد. شکل ۱ مثالی از یک نورون را به صورت انتزاعی نشان می‌دهد.



شکل ۱- عملکرد محاسباتی یک نورون

[1] Multi-layer Perceptron
[2] Multiply And Accumulator
[3] edge

پیش نیازهای انجام تمرین

آشنایی با وریلاگ و یک شبیه‌ساز مانند Modelsim

مراحل انجام تمرین

با توجه به شکل ۲، ماژولی به نام PE طراحی و پیاده‌سازی کنید که به صورت زیر باشد:

- دو عدد حافظه SRAM برای ذخیره سازی وزن‌ها وجود دارد که به صورت پینگ پنگی عمل می‌کنند. فرض کنید یک PE عملیات دو نورون را به ترتیب انجام دهد؛ بنابراین ابتدا وزن‌های نورون اول در SRAM شماره یک ذخیره می‌شوند، پس از اتمام ذخیره سازی همه وزن‌های نورون اول، هنگامی که PE مشغول محاسبات نورون اول است، وزن‌های نورون بعد می‌توانند به صورت موازی در SRAM شماره دو ذخیره شوند و برعکس هنگام محاسبات نورون دوم، وزنهای نورون اول در SRAM اول ذخیره می شوند.
- یک واحد ضرب کننده و یک واحد جمع کننده که عملیات MAC را انجام می‌دهند، در یک PE وجود خواهند داشت. در هر زمان، یک وزن و یک ورودی وارد ماژول MAC شده و نتیجه عملیات در سیکل بعد حاضر می‌شود.
- یک ماژول ReLu وجود دارد به طوری که خروجی ماژول MAC ورودی ReLu خواهد بود. این ماژول ورودی‌های مثبت را عبور داده و بقیه را صفر می‌کند.
- یک ماژول Quantizer نیز وجود دارد. این ماژول پس از ReLu قرار گرفته و ورودی ممیز اعشاری ثابت را کوانتایز می‌کند. به عنوان مثال، یک ورودی ۱۶ بیتی با ۱۴ بیت اعشار، به یک عدد ۸ بیتی با ۷ بیت اعشار کوانتایز خواهد شد. برای این کار، ۷ بیت پر ارزش از ۱۴ بیت اعشار جدا شده، و از بخش صحیح عدد نیز (۲ بیت)، ۱ بیت پایین آن جدا می‌شود.
- این واحد پردازشی، شامل دو سیگنال ورودی است که در واقع به ترتیب مربوط به مقدار وزن و مقدار ورودی یک نورون هستند.
- مقدار حاصل جمع نهایی نیز یک عدد است.
- ورودی‌ها حافظه ای برای ذخیره سازی ندارند.
- عملیات خواندن از SRAM، ضرب و جمع، ReLu و Quantize در ۵ مرحله و به صورت پایپلاین انجام می‌شوند.

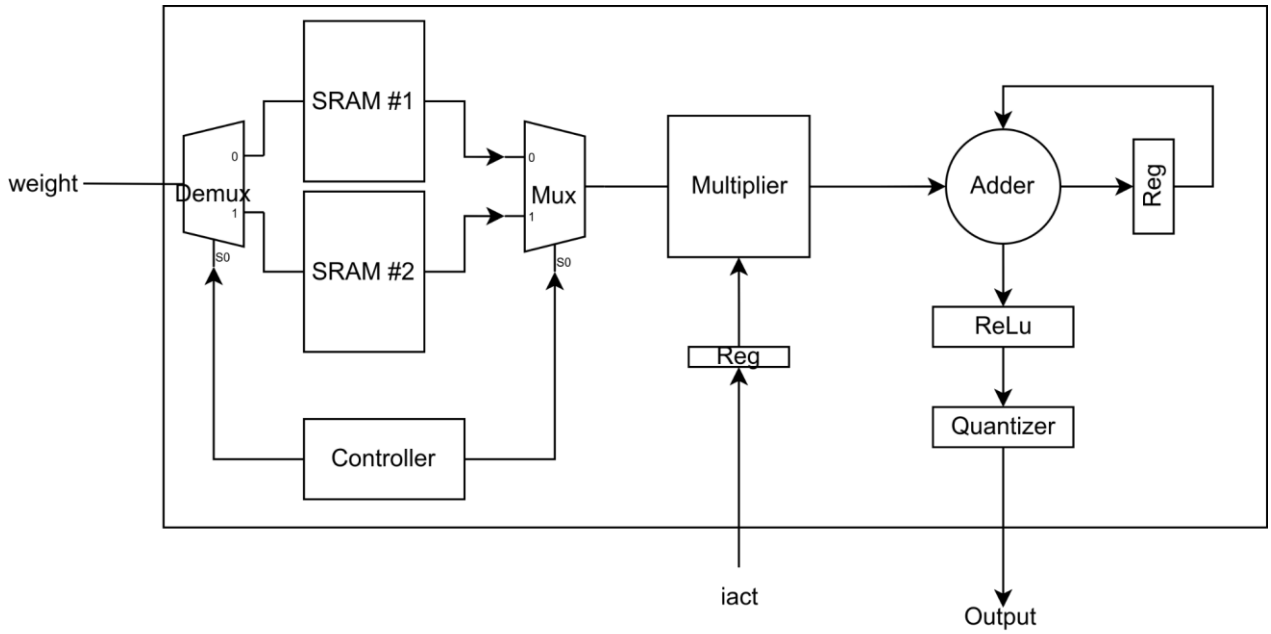
مرحله ۱: طراحی datapath و controller

پیاده‌سازی هر ماژول سخت‌افزاری، ابتدا نیاز به طراحی یک datapath و controller مناسب دارد. هدف این مرحله از تمرین این است که شما برای یک عنصر پردازشی، که عملیات MAC را انجام خواهد داد، طراحی مناسب داشته باشید. با توجه به توضیحات داده شده، در مرحله اول طراحی را انجام دهید. شکل ۲ به صورت کلی datapath اصلی این ماژول را نشان می‌دهد. جزئیات بیشتر را در صورت نیاز به آن اضافه کنید. همچنین اگر به نظر شما قابلیت بهبود دارد، می‌توانید این کار را انجام دهید.

مرحله ۲: پیاده‌سازی

در این مرحله، با توجه طراحی‌ای که در مرحله قبل انجام داده‌اید، PE را با وریلاگ پیاده‌سازی کنید. ماژول‌های مورد نیاز به صورت زیر هستند:

- ماژول حافظه SRAM: هر یک از عملیات نوشتن و خواندن این ماژول یک سیکل طول می‌کشد. یک پورت آدرس و داده جدا برای هر یک از عملیات خواندن و نوشتن در نظر بگیرید. همچنین هر یک سیگنال enable خود را دارند.
- ماژول ضرب کننده: این ماژول در یک سیکل عمل ضرب را انجام می‌دهد.
- ماژول جمع کننده: این ماژول مستقل از کلاک و به صورت ترتیبی عملیات جمع را انجام می‌دهد.



شکل ۲- شمای واحد پردازشی

در نهایت، شما باید با یک PE، خروجی دو نورون را با یک بردار ورودی یکسان تولید کنید. این دو خروجی به ترتیب تولید می‌شوند.

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.
- فایل گزارش به صورت PDF و doc باشد. کد خود را نیز آپلود نمایید.
- تمرین را با فرمت YourName_StudentNo_EAI3.rar آپلود کنید.
- گروه ها حداکثر دو نفر هستند.
- 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 است) از نمره شما کسر می‌شود.
- در صورت مشاهده تشابه زیاد کدها و گزارش، نمره -۱۰۰ برای هر دو گروه اعمال خواهد شد.
- تمرین تحویل حضوری دارد که زمان آن بعدا اعلام خواهد شد.