پیاده سازی تبدیل گفتار به متن بر روی میکروکنترلر

دانشجو: سید محمد حسینی

استاد راهنما: دکتر مهدی کمال

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

خلاصه

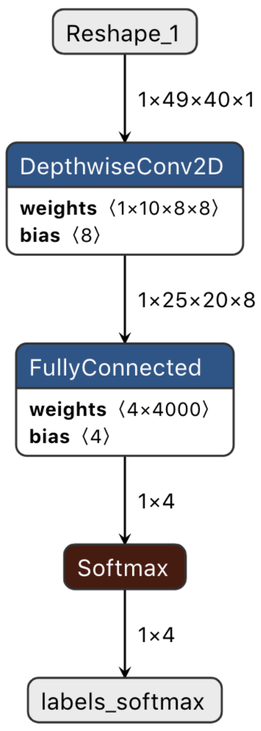
دستیاران صوتی در حال حاضر یکی از موضوعات بسیار داغ و جذاب جهان می‌باشند. این سیستم‌های هوش مصنوعی، با دریافت گفتار انسان و تشخیص آن می‌توانند به عنوان یک راهنما و دستیار در کنار انسان بوده و رنج وسیعی از نیازهای انسان را برطرف کنند.

این دستیاران با دریافت صدای انسان و ارسال آن به سرورهای ابری شرکت سازنده سعی در تبدیل گفتار به متن و سپس پردازش زبان‌های طبیعی آن دارند. این دستگاه‌ها برای دریافت صدا همیشه باید آماده باشند و هرگونه صدا را به سرور ابری شرکت ارسال کنند که این عمل از نظر توان مصرفی بهینه نیست و سیستم توان بسیاری مصرف می‌کند. از سمتی دیگر، دریافت تمام اصوات موجود در محیط و ارسال آن‌ها به سرور ابری باعث ایجاد مشکلات بسیاری در جهت حفظ حریم خصوصی کاربران می‌شود.

در این پروژه با استفاده از ویژگی‌های جدید معرفی شده توسط شرکت‌های بزرگ، سیستم تشخیص کلمات داغ بر روی میکروکنترلر پیاده سازی، تست شده و نتیجه نهایی دریافت شده است. این سیستم از توان مصرفی بسیار کمی برخودار است که باعث می‌شود که بسیار اقتصادی باشد و به دلیل آفلاین بودن حریم خصوصی کاربر را حفظ می‌کند.

ساختار

سیستم تشخیص کلمات با استفاده از میکروکنترلر از ۲ بخش کلی تشکیل شده است. بخش اول مدل یادگرفته شده بر روی یک کامپیوتر قدرتمند است و تبدیل مدل یادگرفته شده به یک مدل بسیار سبک و بخش دوم آن پیاده سازی بر روی میکروکنترلر که یک کامپیوتر بسیار ساده و ضعیف می‌باشد.



شبکه عصبی استفاده در این پروژه با توجه به قیودی که داریم یک شبکه گرافی کوچک می‌باشد. این مدل شامل یک لایه Convolutional همراه با یک لایه Fully Connected و در ادامه یک لایه Softmax برای فعال سازی خروجی ها میباشد. در شکل زیر لایه‌ی convolutional با عنوان DepthwiseConv2D قرار داده شده است.

بعد از پیاده سازی نرم‌افزاری سیستم با استفاده از یک کامپیوتر قدرتمند نیاز است که مدل آموزش داده شده را به یک مدل با حجم بسیار پایین تبدیل کند. این تبدیل با استفاده از ویژگی کتابخانه تنسورفلو به نام TFLite صورت می‌گیرد.

برای پیاده سازی یک شبکه‌ی عصبی بر روی میکروکنترلر نیاز است که معماری نرم‌افزاری - سخت‌افزاری داشته باشیم و سیستم را به صورت ماژولار پیکربندی کنیم. به همین منظور ساختار زیر در پروژه استفاده شده است:

Audio Provider : جمع آوری داده های صوتی محیطی با استفاده از سنسور مورد نظر (داده‌های ورودی)

Feature Provider : پیش پردازش داده‌های ورودی و استخراج ویژگی‌های مورد نیاز شبکه عصبی

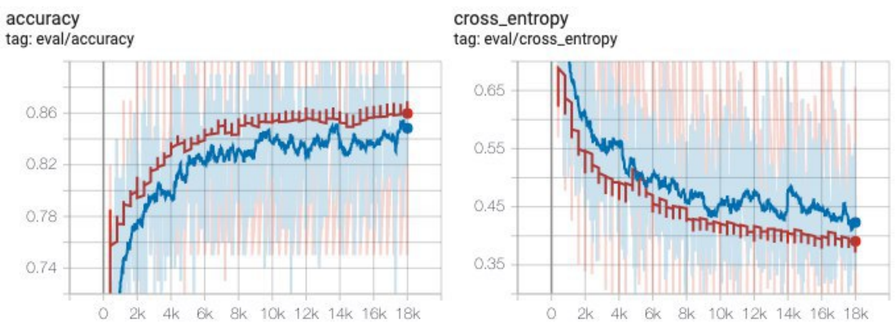
TFLite Interpreter : اجرای شبکه بر روی ویژگی های پردازش شده

Command Recognizer : پس‌پردازش و بدست آوردن خروجی و مشخص کردن دقت‌ها

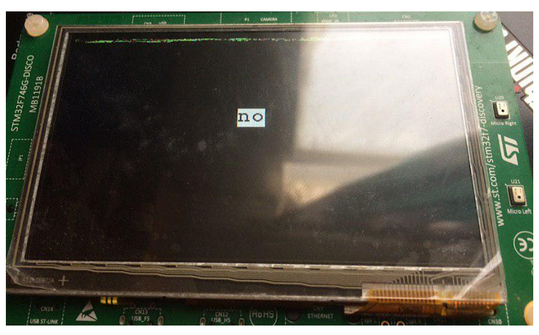
Command Responder : استفاده از نتایج و برای تصمیم گیری

نتایج

دقت نهایی شبکه برای کلمات مورد نظر به ۸۸ درصد رسید. این میزان دقت بهترین ساختار شبکه‌ای بود که با قیود سازگار بود و امکان قرار گیری آن بر روی میکرکنترلر وجود داشت. مدل‌های مختلف دیگری نیز طراحی و آزمایش شد. به عنوان مثال یک لایه تمام اتصال دیگر به شبکه اضافه شد، دقت نهایی آن ۸۹ درصد بود اما حجم خروجی آن ۱۰۰ کیلوبایت بود که در مقایسه با حجم خروجی ۱۸ کیلوبایتی مدل اصلی خیلی تفاوت داشتند. همچنین میزان فلش لازم بر روی میکروکنترلر وجود نداشت تا داده‌ها بر روی آن قرار گیرد.



بعد از تبدیل خروجی به فایل مورد نظر و قرار گیری مدل بر روی سخت‌افزار، با استفاده از میکروفن‌های موجود بر روی بورد مورد استفاده، صدا دریافت و بعد از تشخیص کلمه مورد نظر توسط سخت‌افزار، نتیجه بر روی نمایشگر نمایش داده می‌شود.



جمع بندی

دانش تنهای سخت‌افزاری یا نرم‌افزاری در دنیای امروز دیگر کافی نیست و باید یک مهندس برق و یا کامپیوتر نسبت به دیگر حوزه‌های موجود و مرتبط با توانایی‌هایش آشنایی و در زمان نیاز مهارت پیاده سازی را داشته باشد.

با اتصال دو سیستم جداگانه شبکه‌ی عصبی که برروی یک سیستم بسیار قوی‌تر آموزش داده شده است به یک سخت افزار بسیار ساده و ضعیف، دری به روی پیاده سازی‌های پیشرفته بر روی سیستم‌های توان پایین برای انجام کارهای سطح بالا گشوده شده است.

با استفاده از این پروژه میتوان روزی در صنعت ایران یک دستیار شخصی کاملاً فارسی داشت که بتوان با نمونه‌های خارجی رقابت کند. در این پروژه ما توانستیم که سیستم‌های مختلف را در کنار هم قرار دهیم و یک سیستم یکپارچه شبکه‌ی عصبی در کنار سخت‌افزار بسیار ساده ایجاد کنیم.

مراجع اصلی

1. Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng. *Deep Speech: Scaling up end-to-end speech recognition.*
2. T. N. Sainath, C. Parada, "Convolutional neural networks for small footprint keyword spotting", *Proc. of Interspeech*, 2015.
3. TensorFlow Lite: TensorFlow's lightweight solution for mobile and embedded devices: **https://www.tensorflow.org/lite/**

شماره پوستر: XX