Generativ AI i inbyggda system

Av Jim Beneke, Tria

Jim Beneke är vice president för Tria Americas, ett Avnet-bolag inom inbyggda system. Han har varit 30 år på Avnet och innan dess på Raytheon och GE Aerospace, med allt från forskning och utveckling till strategi och marknadsföring. Han har en BSEE från Bucknell University och en MSEE från Villanova University.

EXPERTARTIKEL

Generativ AI gör det möjligt att använda chattbottar i kundtjänst och smarta högtalare i hemmet. Tekniken med taligenkänning är bara i sin linda men den är ändå på väg att ta steget in i robotiken, där den bidrar till att utveckla algoritmer som bättre styr rörelser och att initiera strategier för att utföra viktiga uppgifter.

Robotar börjar användas i områden där de inte bara interagerar med operatörer utan också direkt med allmänheten. Generativ AI kan göra enorm skillnad för användbarheten genom att tillhandahålla talstyrning och återkoppling. En mobil service-robot på ett hotell eller sjukhus kan exempelvis visa människor vägen dit de ska, eller leverera måltider. Inbyggd tal-till-tal-teknik gör det möjligt för kunder att ställa frågor och få korrekta svar. På samma sätt kan hjälprobotar se till att personer med synnedsättning hittar vägen.

I INDUSTRIELLA TILLÄMPNINGAR, som svetsning och montering, kan roboten lyda röstkommandon och signalera att den förstått dem. Kommandona kan till exempel instruera roboten att flytta en tung del, utföra svetsning och montering, och därefter flytta delen vidare till nästa arbetsstation.

I medicinska sammanhang kan en robot förse en läkare med det instrument som behövs – utan att man behöver bryta sterila rutiner genom att röra vid en skärm eller ett tangentbord.

Många av dagens tal-till-talsystem för konsumentprodukter använder molnet för att leverera sina tjänster. För robotapplikationer går det ofta inte att ha den tidsfördröjning som detta innebär. Dessutom kan industriella och jordbruksrelaterade verksamheter ligga långt ifrån en bredbandsanslutning. Dessa situationer kräver att mycket kraftfulla AI-modeller körs på lokala inbyggnadsplattformar.

TIDIGARE HAR DETTA FÖRKNIPPATS med dyr hårdvara och hög energiförbrukning. Så är inte längre fallet. Med hjälp av den moderna applikationsprocessorn i.MX95 från NXP utvecklade Tria ett system som visar hur generativt tal-till-tal-AI kan porteras till en strömsnål och billig hårdvaruplattform.

Applikationsprocessorn i.MX95 kombinerar ett flerkärnigt Arm-kluster med integrerad grafikprocessor och AI-acceleration baserad på NXP:s neuronprocessor eIQ Neutron, tillsammans med en rad snabba I/O- och minneskontrollers.

Med i.MX95 kan en robot röstkommunicera. Nästa steg blir att hantera rörelser och objekt, navigera, fatta beslut och resonera.

ADOBE STOCK

När man implementerar AI i en inbyggnadstillämpning är det viktigt att välja modeller som erbjuder den bästa avvägningen mellan strömförbrukning, minnesanvändning och noggrannhet. I princip skulle en generativ AI-modell kunna användas hela vägen, end-to-end, men i många fall är detta onödigt. Trias ingenjörer har testat en rad olika optioner för de olika delarna i tal-till-tal-kedjan.

DEN FÖRSTA LÄNKEN är att detektera att en människa gett ett kommando. Detta bör tilldelas en algoritm eller modell som är optimerad för låg energiförbrukning eftersom den måste köras ofta för att roboten inte ska missa viktiga kommandon. Den enklaste algoritmen för detta är ljudstyrkedetektering. Metoden jämför signalen från mikrofonen med nivån på bakgrunden. Även om metoden har extremt låg effektförbrukning har den en oacceptabelt hög andel falska positiva resultat. Däremot erbjuder en Silero-modell för röstdetektering – den bygger på ett vikningsnät (CNN, convolutional neural net) – med hög kvalitet, låg effektförbrukning och liten overhead.

PÅ SAMMA SÄTT fann Trias ingenjörsteam att modellen Piper ger utmärkt prestanda för text-till-tal i förhållande till sin storlek, liksom processor- och minnesanvändning. Det är i länken mellan dessa två steg som ge-

kan röststyra robotar

ADOBE STOCK

En kedja algoritmer avgör att nån pratar, att den ger ett kommando och vad kommandot betyder. Allt kan köras på en i.MX95 – plus kedjan tillbaka, svaret till användaren.

nerativ AI kommer mest till sin rätt. Tekniken som ligger till grund för många av de generativa AI-verktyg som används idag utvecklades för att hantera naturligt språk. En stor språkmodell (LLM) drar nytta av statistiska egenskaper hos mänskligt tal och skrift. Ord och fraser bryts ned i symboler (tokens) som avbildas på ett flerdimensionellt vektorrum (embeddings) på ett sätt som gör att symboler med liknande betydelse hamnar nära varandra. Det är bland annat därför dessa modeller kan hjälpa till att översätta mellan språk.

EN LLM KOMBINERAR VEKTORERNA med neuronnät med en struktur kallad Transformer som använder konceptet attention för att hitta samband mellan tokens som hjälper AI att generera sammanhängande utdata. En stor fördel med träningsprocessen är att den mest beräknings- och datakrävande fasen, känd som förträning (pretraining), inte behöver data som är etiketterad (annoterad). Träningsprocessen låter modellen själv upptäcka samband mellan ord. En andra fas, känd som finjustering, är minst lika viktig. Där används etiketterade data för att optimera den förtränade modellen för en specifik uppgift. För en modell som OpenAI:s Whisper (som är öppen källkod) är denna uppgift att ta diktamen från naturligt tal till text.

Tränad som den är på mer än en halv miljon timmar av flerspråkigt tal i en korpus som representerar många olika typer av uppgifter, är Whisper robust mot brus och dialekter, och kan hantera många typer av fackspråk. Dess relativt lilla storlek, i kombination med viss ytterligare prestanda- och minnesoptimering, gör det möjligt att köra Whisper på inbyggnadsprocessorer.

För applikationen tal-till-tal använde Trias team diskreta tal för att minska modellens beräkningsbörda. AI-modeller som tränas i molnet implementeras vanligtvis i flyttalsaritmetik. Men aritmetikenheterna i processorer som i.MX95 har pipelines som arbetar med heltal.

GENOM ATT KVANTISERA flyttalsparametrarna till 8-bitars heltal (int8) är det möjligt att uppnå dramatiska hastighetsökningar och besparingar i både minnesanvändning och bandbredd, vilket i sin tur minskar energiförbrukningen. Kvantiseringen till int8 betydde att beräkningstiden minskade från 10 sekunder till 1,2 sekunder. För att passa de korta kommandon som förväntas i robotapplikationer minskade teamet dessutom längden på ljudkontexten från 30 sekunder till mindre än två sekunder.

Att bestämma betydelsen av den text som Whisper producerar är en mer komplex uppgift och kräver en större modell anpassad för just det. Stora språkmodeller som kan förstå text tillräckligt bra för att omvandla den till kommandon för en robot kan behöva en miljard eller fler neuronnätsparametrar, även om det är möjligt att minska storleken genom finjustering. För detta tal-till-tal-projekt utvärderade Tria de öppna modellerna Qwen och Llama3, med start på versioner med en miljard parametrar. En central avvägning är hur många tokens en sådan modell kan generera per sekund. Till exempel arbetar Qwenversionen med en halv miljard parametrar på en plattform som i.MX. mer än dubbelt så snabbt som versionen med 1 miljard parametrar,

EN MODELL med en halv miljard parametrar kan ge rimlig funktionalitet när den kombineras med målinriktad finjustering. Du kan till exempel optimera modellen för just de typer av kommandon och svar som en mobil robot förväntas kunna hantera. Utvecklaren kan använda en serverbaserad LLM för att generera en stor del av den etiketterade datan syntetiskt. Detta sparar mycket tid jämfört med manuell generering och etikettering.

För att underlätta integrationen på det Yocto-baserade målsystemet valde teamet en arkitektur byggd kring en tillståndsmaskin, med en MQTT-broker som användes för att vidarebefordra meddelanden mellan modeller och systemkomponenter såsom kamerainmatning och en 3D-avatar implementerad med hjälp av GPU:n. För att säkerställa pålitlig drift körs en watchdog-tråd på processorn som kontrollerar om dikteringen har slutförts inom en angiven tid, och genererar frasen ”kan du upprepa?” om den inte gör det.

GENERATIV AI FÖR TAL-TILL-TAL är bara början. Mer avancerade språkmodeller, multimodala, används nu i forskningsprojekt för att lära robotar att hantera rörelser och objekt. Utvecklare använder även förstärkningsinlärning i kombination med multimodala modeller för att övervinna begränsningarna hos traditionella algoritmer för modellprediktiv styrning. Andra så kallade foundation-modeller med fokus på resonemangsförmåga kommer att göra det möjligt för robotar att navigera utan att förlita sig på kartor, fatta autonoma beslut och sätta ihop strategier för att utföra en uppgift utifrån befintliga regler på lägre nivå.

YTTERLIGARE OPTIMERING av dessa modeller kommer att göra det möjligt att köra dem på kommande energisnåla inbyggnadsplattformar. Fram till dess har robotutvecklaren redan idag tillgång till metoder som gör det möjligt att tala om för en robot vad den ska göra med röstkommandon – och demonstrera att den har förstått uppgiften. ■