

Populärvetenskaplig sammanfattning

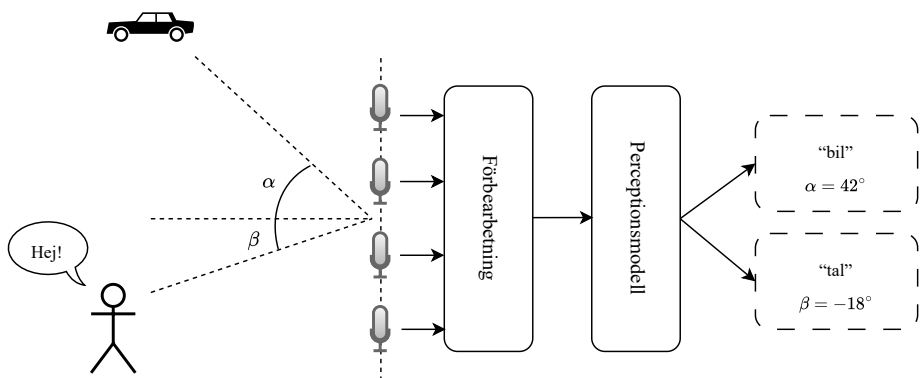
Perception är en viktig komponent i system baserade på artificiell intelligens (AI) som möjliggör bearbetning av signaler från olika sensorer. AI-systemets *sinnen* utgörs av sensorer som omvandlar fysikaliska fenomen till digitala signaler, till exempel kameror, mikrofoner och antenner. Perceptionen består av att tolka dessa signaler genom att extrahera relevant information, så som det semantiska innehållet i kamerabilder eller orden i en talinspelning från mikrofoner. Beroende på systemets användningsområde kan informationen användas på olika sätt. Om systemet används för att styra en autonom robot, kan perceptionen exempelvis användas för att möjliggöra navigering i omgivningen. Perceptiva AI-system kan också användas för att assistera människors beslutsfattande, vilket är vanligt förekommande i bland annat förarassistanssystem.

Numera konstrueras de flesta perceptiva metoder med hjälp av maskininlärning, som bygger på statistiska *modeller* som kan lära sig representationer av signaler genom att träna på stora mängder data. I denna avhandling introduceras flera nya maskininlärningsmetoder, med fokus på bearbetning av ljudsignaler. Dessa metoder innefattar träning av modeller för att klassificera olika typer av ljud från mikrofoninspelningar.

I vissa scenarier kan det vara nödvändigt att även kunna lokalisera ljudkällor i systemets omgivning. Ljudlokalisering har flera viktiga tillämpningar inom bland annat robotik och navigering, och det finns även militära tillämpningar. För att åstadkomma noggrann lokalisering är det nödvändigt att använda en uppsättning med flera mikrofoner, precis som att vår egen förmåga att lokalisera ljud till stor del är beroende av att vi har två öron. Att träna en modell för lokalisering innebär således att behandla en uppsättning signaler och extrahera spatial information som gör det möjligt att uppskatta ljudkällans position.

Det finns ett flertal olika sätt att utvärdera prestandan för perceptiva maskininlärningsmodeller som bör beaktas i designprocessen. Ett viktigt kriterium är att slutledningsförmågan är tillräckligt bra för att systemet ska kunna användas i verklighetstroga användningsområden. Därför är det nödvändigt att träna modellen på en datamängd som innehåller även svåra och ovanliga fall. För en taligenkänningsmodell kan detta innebära exempelvis inspelningar med otydligt uttal eller mycket bakgrundsbrus.

För att ett AI-system skall vara användbart i realistiska scenarier, är det även nödvändigt att kunna implementera det på hårdvara med begränsad beräkningskapacitet. De senaste åren har det emellertid funnits en trend där dessa system blivit mer och mer beräkningsintensiva, vilket till stor del beror på att antalet modellparametrar ökat. Ett viktigt fokusområde i avhandlingen är därför att designa och utvärdera metoder även utifrån ett komplexitetsperspektiv.



Konceptuell överblick av ett perceptionssystem som samtidigt klassificerar och lokaliserar flera ljudhändelser. Inspelningarna från mikrofonerna förbehandlas och bearbetas därefter av en perceptiv modell som klassificerar ljudhändelserna och uppskattar deras position. I denna avhandling introducerar vi inlärningsbaserade metoder för att lösa denna typ av problem.

Illustrationen ovan visar ett förenklat exempel på ett av problemen som behandlas i avhandlingen. Problemet består i att träna en modell som kan detektera, klassificera, lokalisera och spåra olika typer av ljud. Det vanligaste sättet att angripa detta problem är att först använda sig av ett förbearbetningssteg som extraherar två typer av ljudrepresentationer: 1) en spektral representation som är bra för att särskilja olika slags ljud och 2) en spatial representation som innehåller information om ljudkällans position. Den spatiala representationen är emellertid ofta inte tillräckligt bra för att kunna hantera överlappande ljud och bakgrundsbrus, vilket kan leda till felaktiga slutledningar. I denna avhandling introducerar vi därför en metod för att extrahera en adaptiv representation utifrån träningsdata. Vi utvärderar denna representation på realistiska ljudinspelningar och demonstrerar bättre lokaliseringsförmåga jämfört med andra spatiala representationer.

Maskininlärningsmodeller har visat sig vara bra på att lösa perceptionsrelaterade problem på grund av deras förmåga att extrahera representationer av verkligheten från träningsdata. Likväl är deras förmåga också beroende på hur modellernas struktur och hur träningen utformas. Hur bör vi designa dessa modeller för att på bästa sätt extrahera relevant information från en datamängd? Ett möjligt svar på denna fråga är att utnyttja de underliggande symmetrierna som bestämmer förhållandet mellan indata och utdata för det specifika problemet. Läs gärna resten av avhandlingen för att få reda på hur symmetrier kan användas för att förbättra perceptiva modellers prestanda för en mängd olika problem.