Popular Summary

Visually unimpaired people often take the ability to perceive the world with their eyes for granted. When visiting a house for the first time we can recognize the size and layout of the hallway, where to put our shoes, where pets and other people in the room are, and so on. We can also make predictions and decisions based on the visual input we receive. A core challenge of computer vision and artificial intelligence is to develop machines capable of the same thing – perceiving the world around them and acting rationally based on what they observe. In this thesis we study different types of artificial visual perception systems, which can be used for example to automatically detect objects in images or understand human poses and motion in videos.

Today's visual perception systems are typically powered by so called deep neural networks, which are inspired by the human brain with its neurons and complex web of connections. While deep networks yield remarkable results in many applications, they are often computationally expensive and time-consuming to use. This can be especially problematic in real-time scenarios such as video surveillance, or in robotics where an agent may have to quickly explore a large and unknown environment. Also, to make a deep network function properly it is first trained on large amounts of data, typically annotated images. Annotation is a tedious process that costs time and money, as it involves humans describing what the data contains, for example by drawing object boundaries in images. Finally, even when a perception system has been trained it may work poorly in circumstances that differ from the training data. For example, if the perception system is mostly trained on images that depict objects from the front it may fail to recognize them from the side.

In this thesis we study and develop *active* methods for visual perception. By focusing a pre-trained perception model on the most relevant aspects of a scene or an image, computational costs can be reduced and/or conditions where the model is inaccurate can be avoided. We also show how similar ideas can be applied when training perception systems, which reduces the effort associated with data annotation. The active visual perception methods we develop are based on reinforcement learning, a trial-and-error approach for discovering desirable behaviour by means of a reward function. For illustration, consider a self-driving car that should drive from a start location to a given destination within a specific time limit. In practice there may exist several paths between the two locations, such as when driving

in a large city. A simple[1] reward function for this task is the negative distance between the destination and the location of the car when the time is over. This implies that the maximum reward is obtained when the car reaches its goal on time. Note that the reward does not specify *how* the car should drive, only *what* its objective is. Thus the car has to try many different strategies to figure out what works and what does not. Reinforcement learning is suitable in scenarios like this, where an agent may have to perform several actions until it knows whether or not it has succeeded.

This thesis explores active visual perception in three different settings. In the first two we propose methods that actively select what parts of a given input or set of inputs to analyze (from which viewpoints to observe a scene, and where to look in an image, respectively) so that a pre-trained perception system performs well, and/or to reduce the amount of computation that is required. In the third setup we develop and study agents which are tasked to refine a given perception model by actively exploring a given scene, such as a floor plan of a house. As these agents move around the scene they are allowed to ask for annotations (training data), which are then used to refine their perception models. The crux is that the agents are allowed to request only a limited amount of training data, so they should be careful regarding which data they select for training. We show in each setting that active visual perception methods trained with reinforcement learning match or outperform alternative approaches, typically at the same or lower computational costs.

[1]The author of this thesis recommends providing also a negative reward for collisions.

Populärvetenskaplig sammanfattning

De av oss som inte lider av någon synskada tar ofta förmågan att se vår omgivning för given. När vi besöker ett hus för första gången kan vi med synen uppfatta hallens storlek och utformning, var vi kan ställa våra skor, var i hallen som husdjur och andra människor befinner sig, och så vidare. Baserat på vad vi ser kan vi dessutom förutsäga saker och ta relevanta beslut. En av de stora utmaningarna och förhoppningarna inom datorseende och artificiell intelligens är att utveckla maskiner som kan göra samma sak – att se världen omkring dem och agera rationellt baserat på vad de ser. Förmågan att se kallas ofta i mer tekniska sammanhang för visuell perception. I denna avhandling studerar vi olika typer av system för artificiell visuell perception, vilka kan användas exempelvis till att automatiskt detektera objekt i bilder eller förstå människors hållningar (poser) och rörelser i videor.

Dagens visuella perceptionssystem drivs oftast av så kallade djupa neuronnät, vilka är inspirerade av den mänskliga hjärnan med sina neuroner och neuronsammankopplingar. Djupa neuronnät ger idag utmärkta resultat i många tillämpningar, men de är ofta beräkningsmässigt dyra och tidskrävande att använda. Detta kan bli särskilt problematiskt i sammanhang som kräver effektiv bearbetning av data (exempelvis videoövervakning), eller inom robotik där en agent snabbt kan behöva utforska en stor och okänd omgivning. För att få ett djupt neuronnät att fungera som det ska behöver det dessutom tränas på stora mängder data, vanligen annoterade bilder. Annotering är en mödosam process som kostar både tid och pengar, eftersom den involverar människor som beskriver vad bilderna föreställer, exempelvis genom rita konturer kring olika objekt för att markera var i bilden de är och vilken form de har. Ett ytterligare problem är att när ett perceptionssystem väl har tränats kan det fungera betydligt sämre om det används under omständigheter som skiljer sig från träningsdatan. Exempelvis kan ett perceptionssystem som mestadels tränats på bilder av objekt framifrån misslyckas att känna igen dem från sidan.

I denna avhandling studerar och utvecklar vi *aktiva* metoder för visuell perception. Genom att fokusera en redan tränad perceptionsmodell på de mest relevanta aspekterna av en scen eller bild kan man minska mängden beräkningar som behöver göras och/eller undvika omständigheter där modellen ger opålitliga resultat. Vi visar även hur liknande idéer kan appliceras när man tränar ett perceptionssystem, vilket kan reducera mängden dataannotering som krävs. Våra aktiva visuella perceptionsmodeller baseras på förstärkningsinlärning, ett slags prova-och-se-metod för att upptäcka önskvärt beteende baserat på en given belönings-

signal. För att illustrera detta koncept kan man föreställa sig exempelvis en självkörande bil vars uppgift är att köra från en startposition till en given målposition inom en viss tidsram. I praktiken kan det finnas flera vägar mellan de två platserna, till exempel om bilen navigerar i en större stad. En enkel[2] belöningssignal för denna uppgift är den negativa distansen mellan målpositionen och bilens position när tiden är över. Detta innebär att den maximala belöningen erhålls när bilen når sitt mål i tid. Notera att belöningssignalen inte specificerar *hur* bilen bör köra, bara *vad* dess ultimata uppgift är (i det här fallet att nå målpositionen inom en viss tid). Således måste bilen utforska flera olika strategier för att lista ut vad som fungerar och vad som inte gör det. Förstärkningsinlärning lämpar sig väl i den här typen av situationer, det vill säga när en agent behöver utföra flera olika handlingar innan den vet om den lyckats eller inte.

Denna avhandling utforskar aktiv visuell perception i tre olika kontexter. I de första två utvecklar vi metoder som aktivt väljer vilka delar av en insignal eller uppsättning insignaler som ska analyseras (från vilka vyer en scen ska betraktas, respektive var man ska titta i en given bild) för att ett på förhand tränat perceptionssystem ska ge mer pålitliga resultat och/eller för att minska mängden utförda bräkningar. I den tredje kontexten utvecklar och studerar vi agenter vars uppgift är att förbättra en given perceptionsmodell genom att aktivt gå runt och utforska en scen, såsom ett våningsplan i ett hus. Under tiden som agenterna utforskar scenen har de också möjlighet att be om annoteringar (träningsdata), vilka sedan används till att förbättra deras perceptionsmodeller. Kruxet är att agenterna bara tillåts be om en begränsad mängd träningsdata, så de bör vara selektiva angående vilken träningsdata de väljer. Vi visar i samtliga kontexter att aktiva visuella perceptionsmetoder som tränats med förstärkningsinlärning matchar eller förbättrar alternativa metoder, och dessutom oftast till samma eller lägre beräkningsmässiga kostnader.

[2]Författaren till denna avhandling rekommenderar att man även ger en negativ belöning för kollisioner.