4 Popular Summary

Recent advances in artificial intelligence have opened the gate for transformative technologies such as self-driving cars, robot collaborators, and virtual reality. However, technologies such as these rely on computers being able to see humans. For example, a self-driving car must be able to track pedestrians as they move to avoid potentially fatal accidents. Likewise, virtual reality applications require capturing human motion to drive the virtual avatar's movement.

Unfortunately, the fundamental task of perceiving humans is difficult for computers. One reason is that humans vary greatly in appearance; another is that humans can perform a wide variety of actions and fast-paced motions. To make matters worse, the camera might not be able to fully see the human subject due to occlusions or poor lighting conditions. Despite these challenges, impressive progress has been made thanks to so-called artificial neural networks. These mathematical models are inspired by how neurons function in biological brains and learn how to solve tasks from labeled data. For perceiving humans, labeled data usually means videos of humans with corresponding motion capture data in the form of 3d body joint locations tracked over time. Unfortunately, collecting this type of data is time-consuming and requires expensive motion capture equipment. In addition, the capture takes place in indoor studios, limiting the kind of actions that can be captured. Together these factors make it both expensive and challenging to create large and diverse datasets of human motion. As a result, the neural networks often make mistakes when given a new video of humans where the motion or camera viewpoint is different from the examples they were trained on.

This thesis presents methods for improving visual perception through three main ideas. First, we study how an agent equipped with a neural network for perceiving humans should move around to better view the subjects. Imagine a drone tasked with capturing a human athlete. It must move to avoid occlusions and observe the subject from viewpoints seen in the training data to get accurate results. We present neural network-based agents capable of intelligently moving to observe the subjects.

Next, we focus on how an agent should continue to train its neural network once deployed. We study this problem for neural networks performing semantic segmentation (labeling each pixel in an image with an object class). Imagine, for example, an autonomous household robot coming to a new home. The robot might encounter many new objects not present in the training data. In those cases, it would be beneficial if the agent could realize what it does not understand and request explanations from its owner. In this thesis, we present an agent that learns to explore virtual 3d houses and ask for labels for objects it does not recognize. To promote efficiency, we give the agent a budget of how much help it can receive so that it only asks for help when it is crucial.

Lastly, we improve the reconstruction of human motion by integrating the laws of physics. Despite being trained on large video datasets of human motion, neural networks tend to make physically implausible mistakes. For example, the network might predict humans sliding along the ground rather than walking or penetrating the floor. We present a method to combine physics simulation with a neural network perception module to make the results physically plausible – without requiring additional training data.

5 Populärvetenskaplig sammanfattning

Den senaste tidens framgångar inom artificiell intelligens har öppnat dörren till många transformativa tekniker, exempelvis självkörande bilar, robotassistenter och virtuell verklighet. Dessa tekniker bygger på att datorer kan "se människor". En självkörande bil måste kunna hålla uppsikt över fotgängare för att undvika allvarliga olyckor. Likaså kräver datorsimulerad verklighet att datorn kan fånga användarens rörelser för att kunna driva avataren i den simulerade verkligheten.

Tyvärr är det mycket svårt för datorer att uppfatta och urskilja människor. En av anledningarna är att vi människor skiljer oss mycket från varandra när det kommer till utseendet. En annan anledning är att vi kan utföra många olika typer av snabba rörelser. Dessutom kan robotens kamera vara utsatt för dåliga ljusförhållanden eller vara så inställd att delar av objektet faller utanför kamerans synfält. Trots dessa utmaningar har artificiella neurala nätverk lett till stora framgångar. Neurala nätverk är matematiska modeller som bygger på hur biologiska hjärnor attackerar problemen för att lösa dem. När det gäller människors rörelser exemplifieras detta oftast i videor med tillhörande "motion capture" data, där man har spelat in människans leder i 3D. Detta är tyvärr en tidskrävande process som kräver en dyr specialutrustning. Dessutom sker inspelningen av exempel oftast i laboratorier inomhus, vilket begränsar urvalet av rörelser som kan komma ifråga. Detta resulterar i att det är både dyrt och svårt att spela in stora och varierade dataset av mänsklig rörelse. Resultatet blir att de artificiella neurala nätverken, som lär sig av denna ofullständiga data, gör misstag när de observerar människor som utför rörelser som inte finns med i träningsdatan eller när de ser människor från en kameravinkel som avviker från den som användes vid inspelningen i laboratoriet.

Denna avhandling presenterar tre metoder för hur neurala nätverk bättre kan uppfatta omgivningen. I den första delen studeras hur en robot, som är utrustad med ett neutralt nätverk designat för att uppfatta en människa, bör röra sig. Föreställ er en drönare vars uppgift är att spela in en idrottsman. Drönaren bör röra sig så att den har full uppsikt och ser den som idrottar ur de vinklar som det neurala nätverket har tränats för. I min avhandling presenterar jag en artificiell agent som lär sig hur den bäst ska röra sig för att uppfatta en människas rörelser.

I den andra delen studeras hur en agent succesivt förbättrar sitt neurala nätverk när den väl har satts i bruk. Föreställ er en hushållsrobot som har kommit till ett nytt hem. Det kan finnas många nya föremål som roboten aldrig tränats att känna igen. Det vore då fördelaktigt om roboten kunde konkludera, att den inte känner dessa föremål och be ägaren om hjälp med identifikationen. Denna del av avhandlingen presenterar en artificiell agent som lär sig att utforska virtuella hus i 3D och ber om hjälp för att känna igen nya föremål. För att roboten ska arbeta så effektivt som möjligt tillåts den endast att ställa ett begränsat antal

frågor. Därigenom tvingas den välja sina frågor väl.

Den sista delen behandlar hur fysikaliska lagar kan tillämpas för att underlätta för ett artificiellt neuralt nätverk att känna igen en människas rörelse. Trots att nätverken har tränats med omfattande dataset av mänskliga rörelser tenderar de ge orealistiska resultat. Nätverken kan exempelvis göra fel som får det att se ut att en människa svävar istället för att gå framåt. De sista artiklarna presenterar en metod som kombinerar en fysiksimulator med ett neuralt nätverk, vilket gör resultaten mer realistiska. Metoden kräver ingen ytterligare träningsdata.